Prevention of the spread of antibiotic resistance via the structural and molecular study of pLS20 conjugative plasmid proteins

Nerea Bernardo

Universitat Autònoma de Barcelona

Prevention of the spread of antibiotic resistance via the structural and molecular study of pLS20 conjugative plasmid proteins

Nerea Bernardo

Universitat Autònoma de Barcelona

Biochemistry and Molecular Biology Department

Prevention of the spread of antibiotic resistance via the structural and molecular study of pLS20 conjugative plasmid proteins

Doctoral thesis presented by Nerea Bernardo in candidacy for the degree of Ph.D in Biochemistry, Molecular Biology and Biomedicine from the Universitat Autònoma de Barcelona

This Ph.D thesis has been carried out at ALBA Synchrotron under the supervision of Dr. Roeland Boer

Dr. Roeland Boer **Nerea Bernardo**

Cerdanyola del Vallès, Barcelona

September 20th, 2020

“Science and everyday life cannot and should not be separated.”

Rosalind Franklin

Acknowledgments

Here it is. The PhD adventure has come to an end. A 3 years-old journey summarized in about a hundred pages that is far from what it has actually meant to me. It has been a journey full of ups and downs, a bigger amount of negative results than positive ones and days that the main wonder was where all of that was taking me. Yet, it has also been a journey of a great professional and personal growth, constant learn and a reacquirement of the excitement of science. I believe that the achievement of the project can't be reached by the single effort of a person that the project has been assigned to. Hereby, I would like to express my gratitude to everyone who has contributed directly or indirectly to this work.

Firstly, to my thesis director Dr. Roeland Boer for giving me the opportunity to work with him and for his constant supervision and guidance. Thanks for the support and the 24/7 (literally, specially during diffracting nights) availability. Also, for always making time for catching up with the project updates and for data treatment that led to the discovery of new protein structures (or to an another p27 structure).

I would also like to thank Isidro Crespo because he was the best guide when I arrived at the lab and also best pioneer to the structural studies of the project. I will always be grateful for his clear and detailed explanations and permanent aid when encountering different problems in the lab.

To Anna Cuppari, thanks to whom the development of one of the chapters of this work has been feasible. Also, for sharing her expertise with me and allowing me to learn some tricks for data treatment among others.

To Dr. Wilfried JJ Meijer and his team in "Centro de Biologia Molecular Severo Ochoa" for the supply of protein clones, all the hypothesis given in the exchanged mails and functional assays who have been a necessary requisite to understand and contextualize the results obtained in this work.

To all the people with whom I have had the opportunity to work or share spare moments in these years: Albert, Laia, Anabel, Ania, Dani, Anna and Isidro, among others. The working environment would not have been the same without you. Also many thanks to MX people: Fernando, Barbara, Xavi, Damià, Judith and Roeland.

Obviously, none of this would have been the same without the support from my parents and friends. A special thanks to Sabia, Arrate, Carol, Juli and JP for always making time to cheering me up with chocolate or the planning of a new travel.

Lastly, but not less importantly, to Marc. For his unlimited patience and the aid even when he had many things going on. You are a reference point in my life.

To finish, Severo Ochoa once said "*Me he dedicado a estudiar la vida y no sé por qué ni para qué*". This is probably how many people doing life sciences feel on a monthly or even weekly basis. I lack the answer for this too, but at least I can say I have enjoyed the ride.

Moltes gràcies. Eskerrik asko guztioi bihotz-bihotzez.

Abstract

Antibiotics have always been considered one of the most important discoveries of the 20th century. This is indeed true, but nowadays we are facing a real issue, and that is the rise of antibiotic resistance. Genetic capacity of bacteria has improved due to mankind's overuse and misuse of antibiotics and, therefore, they have developed antibiotic resistance genes during these last decades. These genes can be spread among bacterial populations via horizontal gene transfer. Conjugation is the horizontal gene transfer route that is predominantly responsible for the spread of antibiotic resistance genes. Conjugative elements can be placed in the bacterial chromosome or they can also be integrated in plasmids, which are extra-chromosomal, autonomously replicating units. Although several aspects of conjugation have been studied thoroughly over the last years, most of these studies concern Gram-negative bacteria. The human gut has favorable conditions for conjugative gene exchange and has been proven to function as an antibiotic resistance genes pool. A remarkable fraction of Gram-positive bacteria in human gut belongs to the phylum Firmicutes. This PhD thesis studies concern the native conjugative plasmid pLS20 of the Gram-positive bacterium *Bacillus subtilis*, which is a Firmicute. Also, pLS20 has biotechnological interest because of its occurrence in *B. subtilis* natto, which is important in food poisoning. This PhD thesis studies concern the native conjugative plasmid pLS20.

The basis of conjugation process is conserved in both Gram-positive and Gram-negative bacteria. The main player of the conjugative system is the relaxase, however there are several other proteins that also have a crucial role in the process. The relaxase, together with its auxiliary proteins, is in charge of generating the T-strand and transfer to the recipient cell. In this work, we have studied some proteins that are in charge of different processes of pLS20 conjugation. Collecting information of the proteins that take part in the conjugative process in one step or another is of great importance in order to be able to develop new methods to stop the antibiotic resistance genes spread.

The conjugative process needs to be strictly regulated and therefore, there are diverse proteins in charge of it at different levels. One of these proteins is Rco$_{\text{pLS20}}$, which works as the inhibitor of the pLS20 conjugation. Rco$_{\text{pLS20}}$ binds to its promoter via a mechanism known as DNA looping, which involves an Rco$_{\text{pLS20}}$ tetramer forming a loop in the promoter region. We have solved the structure of Rco$_{\text{pLS20}}$ tetramerization domain and have realized that it is not a common folding arrangement in bacteria. Surprisingly, it turned out to have a very similar fold to that of the human p53 oncogen family despite a very low sequence homology. This suggests a structural similarity between the transcription regulation between these two species and brings up the possibility of this fold being ubiquitous.

Another protein in charge of the regulation is Rap$_{\text{pLS20}}$, which works together with its cognate peptide Phr*$_{\text{pLS20}}$. Rap$_{\text{pLS20}}$ activates the expression of conjugative genes by direct binding to Rco$_{\text{pLS20}}$. However, the main determinant of the regulatory state is Phr*$_{\text{pLS20}}$: when it binds to Rap$_{\text{pLS20}}$, Rco$_{\text{pLS20}}$ is released and can again bind to DNA and thus, conjugation is inhibited. We have characterized the binding between Rap$_{\text{pLS20}}$ and Rco$_{\text{pLS20}}$ at different stoichiometries via size

exclusion chromatography. Also, we have evaluated and studied the possibility of cross-regulation of Rap$_{pLS20}$ with other Rap systems, studying binding between Rap$_{pLS20}$ and Phr*F-based peptides, which is the cognate peptide of RapF.

Furthermore, we have worked with Reg$_{576}$, which is also a regulatory protein but takes part in the establishment of the genes once they have been transferred to the receptor cell. In this thesis, we have determined its binding region in DNA and obtained distinct Reg$_{576}$ apo structures crystalizing at different space groups under different conditions, revealing it is a protein that forms crystals with relative ease. One of the structures obtained has been evaluated as a potential model for Reg$_{576}$ disposition when bound to DNA.

Finally, we studied P34$_{pLS20}$ protein, which is predicted to be involved in cell recognition and adhesion. We revealed that it contains a thioester domain, known as TED, which belongs to Class II as discovered by solving its structure. The upper lobe corresponds to the canonical Class I TED fold while the bottom lobe has an immunoglobulin-like fold. Furthermore, we have also structurally characterized a mutant form of P34$_{pLS20}$, P34$_{pLS20}$C68S, with a mutation introduced in the thioester bond forming cysteine. The mutant overall structure is identical, but the bond formation is impaired. By functional assays, a considerable decrease in conjugation yield was observed with the mutant protein.

By understanding the risk of antibiotic resistance spread, we realize about the relevance of studying relaxases and conjugation elements in a functional, biochemical and structural level. Hence, with the information gathered in this work we are now closer to being able to gain insights into the mentioned process and devise rational approaches to reduce the development and spread of antibiotic resistance and, thus, avoid a critical situation in therapy, a return to a pre-antibiotic era.

Resumen

Los antibióticos siempre han sido considerados uno de los descubrimientos más importantes del siglo XX. No cabe duda de que así sea, pero actualmente estamos haciendo frente a un grave problema: el incremento de la resistencia a antibióticos. La capacidad genética de las bacterias ha mejorado debido al exceso e incorrecto use de antibióticos, por lo tanto, durante estas últimas décadas han desarrollado genes que les dotan de resistencia a antibióticos. Estos genes pueden ser propagados entre poblaciones bacterianas mediante transferencia horizontal de genes. La conjugación es la principal ruta de transferencia de material genético que causa la propagación de genes de resistencia a antibióticos. Los elementos conjugativos pueden estar tanto en el cromosoma bacteriano como en plásmidos, los cuales son elementos extra-cromosomales que se replican de manera autónoma. A pesar de que muchos aspectos de la conjugación han sido estudiados durante los últimos años, la mayoría de los estudios están relacionados con bacterias Gram-negativas. El intestino humano posee condiciones favorables para el intercambio de genes mediante conjugación y ha sido demostrado que funciona como una reserva genética de resistencia a antibióticos. Una fracción destacable de bacterias Gram-positivas del intestino humano pertenece al filo Firmicutes. En este trabajo, se ha estudiado el plásmido conjugativo pLS20, de la bacteria Gram-positiva *Bacillus subtilis*, siendo este un Firmicute. Además, pLS20 tiene interés biotecnológico debido a su existencia en *B. subtilis* natto, el cual es importante en intoxicación alimentaria. Esta tesis doctoral trata del plásmido conjugativo pLS20.

La base del proceso conjugativo está conservado en las bacterias Gram-positivas y Gram-negativas. El elemento principal del sistema conjugativo es la relaxasa. Sin embargo, hay muchas otras proteínas que también tienen un papel fundamental en el proceso. La relaxasa, junto con sus proteínas auxiliares, es la encargada de generar una hebra T y de su transferencia a la célula receptora. En este trabajo, hemos estudiado proteínas responsables de distintos procesos en la conjugación de pLS20. Recabar información acerca de las proteínas que participan en este proceso de una manera u otra es de vital importancia si queremos desarrollar nuevos métodos para parar la propagación de los genes de resistencia a antibióticos.

La conjugación ha de estar estrictamente regulada y para ello, hay varias proteínas encargadas de la regulación a distintos niveles. Una de ellas es Rco$_{\text{pLS20}}$, la cual funciona como inhibidora de la conjugación del plásmido pLS20. Rco$_{\text{pLS20}}$ se une a su promotor mediante un mecanismo conocido como “DNA looping”, que consiste en la formación un tetrámero de Rco$_{\text{pLS20}}$ generando un lazo en la región del promotor. Hemos resuelto la estructura del dominio de tetramerización de Rco$_{\text{pLS20}}$ y hemos visto que no se trata de un dominio común en bacterias. Sorprendentemente, resultó compartir parecido con el dominio de tetramerización de la familia de oncogenes p53 a pesar de la baja homología en sus secuencias. Esto hace surgir nuevas preguntas debido a la inexistente unión evolutiva entre estas dos especies y trae la posibilidad de que este tipo de plegamiento sea ubicuo.

Otra de las proteínas encargadas de la regulación es Rap$_{\text{pLS20}}$, la cual trabaja en conjunto con su péptido cognado Phr*$_{\text{pLS20}}$. Rap$_{\text{pLS20}}$ activa la expresión de genes conjugativos mediante su

unión a $\mathrm{Rco_{pLS20}}$. No obstante, el mayor determinante del sistema es $\mathrm{Phr^*_{pLS20}}$, ya que debido a su unión a $\mathrm{Rap_{pLS20}}$, $\mathrm{Rco_{pLS20}}$ queda liberada y se une al DNA e inhibir la conjugación. Hemos caracterizado la unión entre $\mathrm{Rap_{pLS20}}$ y $\mathrm{Rco_{pLS20}}$ a distintas estequiometrías utilizando cromatografía de exclusión molecular. Además, hemos evaluado y estudiado la posibilidad de regulación cruzada entre $\mathrm{Rap_{pLS20}}$ y otros sistemas de Rap, mediante ensayos de unión de $\mathrm{Rap_{pLS20}}$ y péptidos basados en Phr*F, el cual es el péptido cognado de RapF.

Por otro lado, hemos trabajado con $\mathrm{Reg_{576}}$, la cual es una proteína reguladora que participa en el establecimiento de genes una vez transferidos a la célula receptora. Hemos determinado su región de unión en el ADN y hemos obtenido varias estructuras de $\mathrm{Reg_{576}}$ en su forma apo en distintos grupos espaciales y bajo diferentes condiciones de cristalización, lo cual revela que es una proteína que forma cristales con relativa facilidad. Una de las estructura obtenidas ha sido evaluada como un modelo potencial de la disposición de $\mathrm{Reg_{576}}$ en la unión al ADN.

Por último, hemos estudiado la $\mathrm{P34_{pLS20}}$, proteína involucrada en reconocimiento y adhesión celular según predicciones. Contiene un dominio tioéster, denominado TED, que pertenece a la Clase II basándonos en su estructura. El lóbulo superior corresponde al plegamiento de Clase I TED canónico mientras que el lóbulo inferior posee un plegamiento del tipo inmunoglobulina. Además, hemos caracterizado estructuralmente una forma mutante de $\mathrm{P34_{pLS20}}$, $\mathrm{P34_{pLS20}C68S}$, el cual tiene una mutación introducida en la cisteína que participa en el enlace tioéster. La estructura global del mutante es parecida a la nativa, aunque la formación del enlace se ve afectado. Mediante ensayos funcionales, un descenso considerable en la rendimiento de la conjugación ha sido observado.

Entendiendo el riesgo de la propagación de la resistencia a antibióticos, nos damos cuenta de la relevancia del estudio de relaxasas y elementos conjugativos de una manera funcional, bioquímica y estructural. Así, con la información que hemos obtenido mediante este trabajo, estamos más cerca de poder entender el proceso conjugativo e idear enfoques racionales para reducir el desarrollo de la propagación de resistencia a antibióticos y por lo tanto, evitar una situación crítica, el retorno a una era sin antibióticos.

Resum

Els antibiòtics sempre han sigut considerats un dels descobriments més importants del segle XX. No hi ha cap dubte de que així sigui, però actualment estem fent front a un greu problema: l'increment de la resistència a antibiòtics. La capacitat genètica dels bacteris ha millorat a causa de l'excés i incorrecte consum del antibiòtics, per tant, aquestes últimes dècades han desenvolupat gens que li doten de resistència a antibiòtics. Aquestes gens poden ser propagats entre les poblacions bacterians via transferència horitzontal de gens. La conjugació és la principal ruta de transferència de material genètic que causa la propagació de gens de resistència a antibiòtics. Els elements conjugatius poden estar al cromosoma bacterià o en plasmidis, els quals són elements extra-cromosomals que es repliquen de manera autònoma. Tot i que molts aspectes de la conjugació han sigut estudiats durant els últims anys, la majoria dels estudis estan relacionats amb els bacteris Gram-positius. L'intestí humà té condicions favorables per l'intercanvi de gens via conjugació i ha sigut demostrat que funciona com una reserva genètica de resistència a antibiòtics. Una fracció destacable de bacteris Gram-positiu de l'intestí humà pertany al filo Firmicutes. En aquest treball, s'ha estudiat el plasmidi conjugatiu pLS20, del bacteri Gram-positiva *Bacillus subtilis*, que és un Firmicute. A més, pLS20 té interès biotecnològic degut a la seva existència en *B. sutiblis* natto, el qual és important en intoxicació alimentària. Aquesta tesi doctoral tracta del plasmidi conjugatiu pLS20.

La base del procés conjugatiu està conservat en els bacteris Gram positius i negatius. L'element principal del sistema conjugatiu és la relaxasa, tot i que hi ha moltes altres proteïnes que també tenen un paper fonamental al procés. La relaxasa, en conjunt amb las seves proteïnes auxiliars, és l'encarregada de generar una *T-strand* i de la seva transferència a la cèl·lula receptora. En aquest treball, hem estudiat els proteïnes responsables de processos en la conjugació de pLS20. Recol·lectar informació sobre les proteïnes que participen en aquest procés d'una manera o altra és molt important si volem desenvolupar nous mètodes per parar la propagació dels gens de resistència a antibiòtics.

La conjugació ha de estar estrictament regulada i per això, hi ha diverses proteïnes encarregades de la regulació a distintes nivells. Una d'aquestes es Rco$_{\text{pLS20}}$, la qual funciona com inhibidora de la conjugació del plasmidi pLS20. Rco$_{\text{pLS20}}$ s'uneix al seu promotor mitjançant un mecanisme conegut com "DNA looping", que consisteix en què un tetràmer de Rco$_{\text{pLS20}}$ formi un llaç en la regió del promotor. Hem resolt l'estructura del domini de tetramerització de Rco$_{\text{pLS20}}$ i hem vist que no es tracta d'un domini comú en bacteris. Sorprenentement, ha resultat compartir semblança amb el domini de tetramerització de la família de oncogens p53 tot i la baixa homologia de las seves seqüències. Això provoca noves preguntes a causa de l'inexistent unió evolutiva entre aquestes dues espècies i obre la possibilitat de què aquest tipus de plegament sigui ubic.

Una altra de les proteïnes encarregades de la regulació és Rap$_{\text{pLS20}}$, la qual treballa en conjunt amb el seu pèptid cognat Phr*$_{\text{pLS20}}$. Rap$_{\text{pLS20}}$ activa l'expressió de gens conjugatius mitjançant la seva unió a Rco$_{\text{pLS20}}$. No obstant, el major determinant del sistema és Phr*$_{\text{pLS20}}$, ja que través de la seva unió a Rap$_{\text{pLS0}}$, Rco$_{\text{pLS20}}$ queda lliure i es pot unir-se al DNA i inhibir la

conjugació. Hem caracteritzat la unió entre Rap_{pLS20} i Rco_{pLS20} a distintes estequiometries utilitzant cromatografia d'exclusió molecular. A més, hem avaluat i estudiat la possibilitat de regulació creuada entre Rap_{pLS20} i altres sistemes de Rap mitjançant assajos d'unió entre Rap_{pLS20} i pèptids basats en Phr*F, el qual és el pèptid cognat de RapF.

Par altra banda, hem treballat amb Reg_{576}, que és una proteïna reguladora que participa en l'establiment de gens una vegada transferits a la cèl·lula receptora. En aquesta tesi, hem determinat la seva regió d'unió al DNA i hem obtingut diverses estructures de Reg_{576} a la seva forma apo en diferents grups espacials i sota diferents condicions de cristal·lització, revelant que és una proteïna que forma cristalls amb relativa facilitat. Una de les estructures obtingudes ha sigut avaluada com un model potencial de la disposició de Reg_{576} en la unió al ADN.

Per últim, hem estudiat la $P34_{pLS20}$, proteïna involucrada en reconeixement y adhesió cel·lular segons prediccions. Conté un domini tioèster, denominat TED, que pertany a la Classe II, en base a en la seva estructura. El lòbul superior correspon al plegament de Classe I TED canònic, mentre que el lòbul inferior té un plegament del tipus immunoglobulina. A més, hem caracteritzat estructuralment una forma mutant de $P34_{pLS20}$, $P34_{pLS20}C68S$, que té una mutació introduïda en la cisteïne que participa al enllaç tioèster. L'estructura global del mutant es semblant a la nativa, tot i que la formació del enllaç es vegi afectat. A través d'assajos funcionals, un descens considerable en el rendiment de la conjugació va ser observat.

Comprenent el risc de la propagació de la resistència a antibiòtics, ens hem adonat de la rellevància del estudi de relaxases i elements conjugatius d'una manera funcional, bioquímica i estructural. Així, amb la informació que hem obtingut en aquest treball, som més a prop de poder entendre el procés conjugatiu i idear enfocaments racionals per reduir el desenvolupament de la propagació de resistència a antibiòtics i per tant, evitar una situació crítica, el retorn a una era sense antibiòtics.

List of Abbreviations

Abbreviation	Meaning
AFM	Atomic force microscopy
AR	Antibiotic resistance
Ard	Alleviation of restriction of DNA
ARG	Antibiotic resistant gene
ASU	Asymmetric unit
AUC	Analytical ultracentrifugation
BLAST	Basic local alignment search tool
C230	Catechol 2,3-dioxygenase
CDC	Centers for Disease Control and Prevention
CE	Common era
CTD	C-terminal domain
D	Translational diffusion coefficient
DBD	DNA Binding domain
DLS	Dynamic light scattering
DR	Direct repeat
ds-DNA	Double-stranded DNA
dso	Double strand Origin
ECDC	European Centre for Disease Prevention
EMSA	Electrophoretic mobility shift assay
FDA	Food and Drug Administration
FL	Full-length
FT	Fourier Transform
GF	Gel filtration
Gram +	Gram positive
Gram -	Gram negative
GRAS	Generally recognized as safe
GTA	Gene transfer agents
GST	Glutathione S-transferase
HDT	Horizontal DNA transfer
Hepes	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid

HGT	Horizontal gene transfer
HTH	Helix-Turn-Helix
ICE	Integrative conjugative element
Ig	Immunoglobulin
IPTG	Isopropyl-β-thiogalactopyranoside
LB	Luria-Bertani
LB/kan	Kanamycin containing Luria-Bertani
LGT	Lateral gene transfer
MAD	Multi- wavelength anomalous diffraction
MBP	Maltose-binding protein
MDR	Multiple drug resistance
MGE	Mobile genetic element
MIR	Multiple isomorphous replacement
MPF	Mating pair formation
MR	Molecular replacement
MRSA	Methicillin-resistant Staphylococcus aureus
MSA	Multiple sequence alignment
MTase	Methyltransferases
M_W	Molecular weight
NTD	N-terminal effector domain
OD	Oligomerization domain
O/N	Overnight
ORF	Open reading frame
oriT	Origin of transfer
PDB	Protein data bank
PEG	Polyethileneglycole
PMF	Protein mass fingerprinting
QS	Quorum-sensing
RBS	Ribosome binding site
RCR	Rolling cycle replication
REase	Restriction endonclease
Rep	Replication initiator
RHH	Ribbon-Helix-Helix
RMSD	Root-mean-square deviation
RSC	Remodeling the structure of chromatin

R_{free}	Free R-factor
R_{work}	Work R-factor
SAD	Single-wavelength anomalous diffraction
SAM	Sterile alpha motif
SAXS	Small angle X-ray scattering
SCOP	Structural classification of proteins
SE	Sedimentation equilibrium
SEC	Size exclusion chromatography
SeMet	Selenomethionine
SNP	Single nucleotide polymorphism
SSB	Single stranded binding
ss-DNA	Single-stranded DNA
sso	Single strand origin
Superdex 75	Superdex 75 Increase 5/150 GL
Superdex 200	Superdex 200 Increase 5/150 GL
SV	Sedimentation velocity
T2SS	Type II secretion system
T4P	Type IV pili
TA	Transactivation domain
TB	Terrific-Broth
TCE	Trichloroethylene
TCM	Traditional chinese medicine
TED	Thioester domain
TID	Transactivation inhibitory domain
TIE	Thioester, isopeptide, ester
TOM	Toluene *ortho*-monooxygenase
TraJ	Transfer genes
TRP	Tetratricopeptide repeat
UV-Vis	Ultraviolet-visible
V_{el}	Elution volume
VRE	Vancomycin-resistant enterococci
WHO	World Health Organization
XRD	X-ray diffraction
Ɛ	Extinction coefficient
Φ	Hydrophobic residue

ζ	Hydrophilic residue

Table of Contents

1. INTRODUCTION

2. MATERIALS & METHODS

Chapter 1: Rco$_{pLS20}$, Inhibitory protein of pLS20 conjugation

Chapter 2: Rap$_{pLS20}$, Phr$_{pLS20}$ and Rco$_{pLS20}$: Regulation of conjugation of pLS20*

Chapter 3: Reg$_{576}$, A regulator of establishment

Chapter 4: $P34_{pLS20}$, Cell adhesion of pLS20

1. INTRODUCTION

1.1. ANTIBIOTIC RESISTANCE: A MAJOR PUBLIC HEALTH PROBLEM

Since the discovery and subsequent development of antibiotics and their derivatives, the number of antibiotic resistant bacteria is constantly incrementing.[1] In the last two decades, the world has witnessed a threatening increase in the number of antibiotic resistant bacterial pathogens. Major global organizations like the World Health Organization (WHO), European Center for Disease Prevention and Control (ECDC) and United States Centers for Disease Control and prevention (CDC) now consider Antibiotic Resistance (AR) as a major and emerging threat to global public health.[2] They consider it to be a crisis with potentially catastrophic consequences as the world would enter a post-antibiotic era.

AR caused 671,689 infections in the year 2015 and contributed to over 33,000 deaths in the European Union.[3] In the United States, based on the CDC report of 2019, more than 2.8 million infections occur and over 35,000 people die each year. The total cost of hospitalization and treatment of patients that had been infected by multidrug-resistant bacteria was estimated to be of at least EUR1.5 billion annually according to the ECDC. Whereas in the United States, based on an estimation done in 2014, $2.2 billion are spent each year.[4] The situation is even more critical in countries with low and middle incomes, as a higher incidence of infectious diseases is observed.[5] Furthermore, AR is estimated to cause around 300 million premature deaths by 2050, with a loss of up to $100 trillion to the global economy.[6] Based on this numbers, the WHO and ECDC identified AR as one of the most important public health problems of the 21st century.

1.2. ANTIBIOTICS

1.2.1. Definition of the term antibiotic

In 1947, S.A. Walkman defined the term antibiotic as “a chemical substance, produced by microorganisms, which has the capacity to inhibit the growth and even destroy bacteria and other microorganisms”.[7] Nowadays, antibiotic can have several meanings: ***i)*** an organic molecule of natural or synthetic origin that either inhibits or kills pathogenic bacteria, ***ii)*** any antimicrobial compound or, ***iii)*** based on Waksman tradition, a microbial substance from a microbial origin.[8]

1.2.2. Antibiotics History

The discovery and large-scale production of antibiotics in the early 20th century was one of the most remarkable achievements in the history of medicine. Thanks to these new drugs, fear of many infectious diseases started to vanish and life expectancy increased considerably. Furthermore, they generated a lot of new knowledge about pathogens. Back in 1900, the most common cause of death was contagious diseases, which nowadays are responsible for only a small percentage of annual deaths. Today, the situation in many countries of the Third World is comparable to that in the 1900, whilst in developed countries it is mostly an issue related to immune-suppressed people infected with multi-resistant pathogens. Before different antibiotics began to be produced and commercialized, mankind was tortured by huge epidemics like smallpox, typhus, tuberculosis, malaria, cholera, leprosy, yellow fever, Spanish flu pandemic and so many more.

An illustrative example of this disastrous situation is that of the plague. This disease is caused by the bacterium *Yersinia pestis* and it is responsible for at least three pandemics in history: ***i)*** In the sixth century, it caused nearly 30-50 million deaths, half the world's population at that time, ***ii)*** a period in the 14th century known as "Black Death" which was the deadliest pandemic recorded in human history. 75-200 million people died in Eurasia and North Africa.[9] ***iii)*** The outbreak between 1895 and 1930, that brought about 12 million victims, most of them in India.[10]

1.2.2.1. The pre-antibiotic era

It is interesting to note that microorganisms have been used to treat several illnesses in ancient times, even though the reason behind their effect remained a mystery. There are several examples of how antibiotics were used in ancient times even if people back then were not aware they were consuming them. For instance, traces of tetracycline have been found in human skeletal remains from ancient Sudanese Nubia from 350-550 CE.[11,12] This fact can only be explained by tetracycline-containing materials in their diet. Another example of ancient antibiotic exposure was discovered after analyzing femoral midshafts of the late Roman period skeletons from Dakhleh Oasis (Egypt) which showed presence of tetracycline in the diet of that time.[13] Consistent with these findings, the rate of infectious diseases in Sudanese Nubia was low and no indications of bone infections were observed in the samples studied from the Dakhleh Oasis.[13,14]

Traces of antibiotics other than tetracyclines are harder to detect as tetracyclines are chelators and are incorporated into the hydroxyapatite mineral portion of the bones and tooth enamel. Evidence of exposure to any other type of antibiotic is based on surviving customs and anecdotal evidences. For example, there are anecdotes of red soils that were used in Jordan for treating skin infections that have led to the discovery of antibiotic producing bacteria.[15] Bacteria isolated from these foils produced actinomycin C2 and actinomycin C3, which are polypeptide antibiotics that bind to a pre-melted DNA conformation that is present in the transcriptional complex, thus it is very unlikely to be detected.[16]

Another example of pre-antibiotic era is the Traditional Chinese Medicine (TCM), which has been used in China for more than four thousand years. Distinct from western medicine, it is based on the use of formulas of herbs that are tailored to individual patient depending on their condition. One of the best-known examples is that of *quinghaose* (*artemisinin*), an anti-malarial drug, which was extracted from *Artemisia* plants in the 1970s.[17] There is a number of other herbs that were used in TCM that have been discovered to have antimicrobial activity, such as *Cortex fraxini*, *Radix arnebia* or *Rhizoma coptidis*.[18,19]

Nevertheless, without the discovery and description of bacteria, a successful development of antibiotics would have been unfeasible. The first important event in microbiology was set by Antonie van Leeuwenhoek in 1677, who was able to visualize and draw bacteria, sperms and erythrocytes using a handcrafted microscope.[20,21]

In the 17th century, the English apothecary John Parkinson described the healing properties of molds and encouraged the use of microorganisms to treat infections in his book *Theatrum botanicum*.[22] Subsequently, and for about two hundred years, there were no remarkable contributions to this topic.

In the second half of the nineteenth century, there were two main questions that scientists were pondering: ***i)*** "Does spontaneous generation exist?" and ***ii)*** "What is the nature of infectious diseases?". After a lot of wrangling in the first topic, French chemist Louis Pasteur finally disproved spontaneous generation by the well-known swan-neck flask (Col de sygnet) experiments in 1859. This way, he showed that contamination was required for microbial growth and marked the end of the two-millennium old theory of spontaneous generation.[23] Regarding the second question, a definite answer was provided by Robert Koch, who successfully showed the connection between a bacterium (in this case, *Bacillus anthracis*) and an illness (anthrax). He visualized and drew moving *B. anthracis* in the blood of infected animals together with the aid of Ferdinand Cohn.[24]

1.2.2.2. The antibiotic era

The beginning of the modern antibiotic era is often linked with two names: Paul Ehrlich and his idea of "magic bullets" and Alexander Fleming and penicillin.

Paul Ehrlich, who is considered the founder of chemotherapy, thought that some chemicals seemed to have antimicrobial properties. He developed an interest in staining tissues with dyes. The reason for such interest was that he believed there were compounds with the ability to selectively target disease-causing bacteria and his idea was to find toxic substances with discriminatory properties for medical purposes against infectious diseases. He aimed to find a drug against syphilis, disease caused by the spirochaete *Trepanoma pallidium,* which was endemic, rampant and almost incurable at that time. Routine therapy against it consisted on inorganic mercury salts, however, the treatment had serious side effects and its efficacy was very low.[25]

In 1905, H. Wolferstan Thomas described aminophenyl arsenic acid (commercially known as atoxyl) as a successful treatment for sleeping sickness[26], which was caused by *Trypanosoma brucei*. As it had had some success in the treatment of this illness, Ehrlich and his colleagues, the bacteriologist Sahachiro Hata and the organic chemist Alfred Bertheim synthesized hundreds of organoarsenic derivatives of the highly toxic drug and tested them *in vivo* in syphilis-infected rabbits. Finally, in 1909, they realized that the compound number 606, which was the arsphenamine (dioxy-diamino-arsenobenzol-dihydrochloride), and afterwards given the trade name of Salvarsan, cured the syphilis-infected rabbits.[27] From November 1910, the pharmaceutical company Hoechst was producing about 12,000-14,000 ampoules of Salvarsan per day that were used for clinical trials and therapy.[28,29,30]

Nonetheless, Salvarsan had side effects like hypersensitivity problems due to arsenic poisoning. Therefore, Ehrlich kept evaluating synthesized organoarsenic derivatives and in 1914, when he tested compound 914, he discovered an improved version of Salvarsan. This compound was neoarsphenamine, which was marked and developed as Neosalvarsan. It had lower arsenical content that resulted in increased solubility and decreased toxicity. Although it had side effects like nausea and vomiting, Salvarsan and the enhanced version Neosalvarsan were both a success and it was the most prescribed drug until penicillin was introduced in the market in the 1940s.[31] Surprisingly, even if these antibiotics are over 100 years old, the exact biochemical mechanism by which these compounds defeated the syphilis spirochete is still unknown and its chemical structure was not determined until 2005.[32]

Ehrlich's contribution to science was not only Salvarsan, but also the systematic screening approach he began to use. Pharmaceutical industry chose to take up this approach and consequently, thousands of drugs were introduced into the market. In 1908, Ehrlich was awarded the Nobel Prize in Medicine together with Ilya Ilyich Mechnikov, who discovered phagocytes, in recognition of their work on immunity.[33]

Another milestone in antibiotic history was set by Alexander Fleming in 1929 when he made the following observation: "It was noticed around a large colony of a contaminating mold the staphylococcus colonies became transparent and were obviously undergoing lysis".[34] Based on this perception, Fleming isolated and grew the fungus in separate culture plates. The fungus was proven to be very effective even when grown at very low concentrations and was less toxic than other disinfectants that were used at that time. This fungus was *Penicillium notatum*, and the antibiotic chemical produced by it, was named penicillin. Even if antibacterial properties of mold had already been seen in ancient times, Fleming was very persistent on this observation but was neither able to elucidate the chemical structure of penicillin nor to produce a significant amount of it. Howard Florey and Ernest Chain, who were a pathologist and a biochemist working at the University of Oxford, discovered the structure of penicillin in 1939, when Fleming was already considering on putting an end to his project. Furthermore, they managed to purify penicillin in quantities that were enough for clinical trials. They were therefore remarkably helpful for scaling up the production of penicillin, which was introduced in therapy in 1941.[35] As a consequence, Florey, Chain and Fleming shared the 1945 Medicine Nobel Prize.

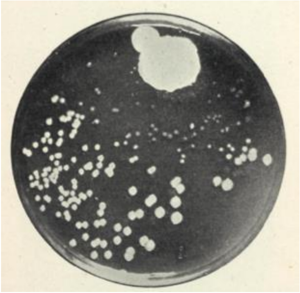


FIGURE 1: **Photographic print of the orginial *Staphylococcus* colony plate showing the *P. notatum* contamination (1928).** *Staphylococcus* undergoing lysis and regular *Staphylococcus* colonies can be seen. *"Image taken from St.Mary Hospital Medical School/Science Photo Library"*

They soon started providing the drug to people suffering from bacterial infections and it turned out to be a great success. They published their clinical findings[36] but were unable to convince pharmaceutical companies in Great Britain to produce penicillin due to World War II commitments. Florey decided to ask United States for assistance. The United States government took over all penicillin production when they entered World War II in 1941. Deep-tank fermentation was used to produce huge quantities of the drug. It was a huge success as by September 1943, the stock of penicillin they had was sufficient to satisfy the demands of the Allied Armed Force.[37]

Penicillin was the first antibiotic that killed Gram-positive (Gram +) bacteria, including the pathogens that caused gonorrhea or syphilis. Aside from that, Fleming's screening method had a remarkable contribution to how antibiotic research became much more efficient. The testing of inhibiting agents in agar plates containing pathogenic bacteria and the subsequent examination of the inhibition zones of these pathogens saved time and money. Furthermore, much fewer resources were needed compared to any kind of testing in animal disease models. Thus, many researchers in academia and industry adopted this method for mass screenings for antibiotic producing microorganisms.[38]

Despite these two discoveries by Ehrlich and Fleming being notorious and well known, the fact that the first hospital use of a drug that we would name an antibiotic today was the so-called *Pyocyanase* remains unknown for many people. Pyocyanase was prepared by Emmerich and Löw in 1899 from *Pseudomonas aeruginosa*[39]. They realized the bacterium and the prepared extracts were active against several pathogenic bacteria. Sadly, the results when they were used for treatment were not persistent and the preparation of the compound itself was fairly toxic for humans, hence treatment by Pyocyanase was stopped.[38]

The discovery of the afore-mentioned antibiotics established the foundations for future drug discovery research. Many other researchers followed the general principles used for their development and it led to many new antibiotics. The years between 1950 and 1970 are known as the golden era for the discovery of new antibiotic classes: tetracyclines, chloramphenicol, amynoglycosides, glycopeptides and so on. Unfortunately, no new classes have been found since then, with the exception of oxazolidinones.[40] Mostly, the approach followed since the end of the golden era has been the modification of existing antibiotics to overcome emerging resistance of pathogens to antibiotics.

1.2.3. Antibiotic Resistance

Antibiotics are definitely one of the most significant medical inventions. Modern medicine was revolutionized with the commercialization and administration of antimicrobial compounds. Sadly, the increased antibiotic resistance (AR) among bacterial pathogens is now threatening one of the most significant medical discoveries of all time.

In order to really comprehend the problem of AR, also known as multidrug resistance (MDR), there are some concepts that need to be understood. Firstly, AR is ancient and it is an outcome of the interaction of several organisms with their environment. Mostly all antimicrobial molecules are spontaneously produced. Therefore, bacteria have developed different strategies to overcome them and to learn to live in their presence. Nevertheless, when we talk about AR, the main target is not these mentioned intrinsic mechanisms, but an "acquired resistance", which means a naturally susceptible microorganism that acquires the ability not to be affected by a given drug.[41] Bacteria can easily acquire resistance by sharing genetic material with one another.

The first sign of AR became apparent soon after the discovery of penicillin. In 1940, it was revealed that an *Escherichia coli* strain could be able to destroy penicillin by producing an enzyme they named penicillanese.[42]

One AR study performed by Rollo *et al.* in 1952 concluded that "Syphilis has now been treated with arsenical for about 40 years without any indications of an increased incidence of arsenic-resistant infections, and this work gives grounds for hoping that the widespread use of pencillin will equally not result in an increasing incidence of infections resistant to pencillin."[43] Therefore, at that time, the outlook of the use of antibiotics was rather idealistic. This mentioned study in 1952 is still relevant for some bacteria like syphilis-causing *Treponema pallidium*.[44] However, other pathogenic bacteria have become resistant to penicillin, semi-synthetic penicillin, cephalosporins, and newer carbapems.[45]

By 1942, four *Staphylococcus aureus* strains were found to resist penicillin in hospitalized patients.[46] During the next few years, infections spread from hospital to communities. By the late 1960s, more than 80% of both community and hospital-acquired strains of *S. aureus* were resistant to penicillin.[47] Thanks to the introduction of second-generation semi-synthetic methicillin, this spread was a bit decelerated. Nevertheless, methicillin-resistant strains soon emerged.[48] In 1967, strains of *S. pneumoniae* also became resistant to penicillin.[49] By 1999, the incidence associated with antibiotic-resistant *Streptococcus pneumonia* was three times of that of 1979.[50] *Enterobacteriaceae* is also a group with high rates of resistance to penicillin, of which several strains are intrinsically aminopenicillin-resistant.[51] Between 1950 and 2001, about 66% of *E. coli* caused human diseases could not be treated by penicillin in the United States.[52]

There are two main mechanisms for bacteria to develop AR that are represented in **FIGURE 2**. On the one hand, preventing the antibiotic from reaching its target may result in AR.[53] For this purpose, bacteria either make use of efflux pumps that remove the antibiotic from the cell, or induce a decreased permeability of the membrane that hampers the entry of the drug into the cell. The second mechanism involves destruction or chemical modification of the antibiotics by enzymes. For instance, β-lactamase destroys an important structural element (the β-lactam ring) of penicillins. Chemical modification can introduce a less expensive modification of the chemical structure of the antibiotic, which impedes it from binding to its target.

On the other hand, modification of the target that the antibiotic works on may provoke the development of AR too.[53] This can be achieved by camouflaging the target by changes in the composition or structure of the target in order to prevent the binding between the drug and the target, expressing proteins that can be used in replacement of the protein inhibited by the antibiotic, or reprograming the target so a different variant of the structure the bacteria needs is produced.[54]

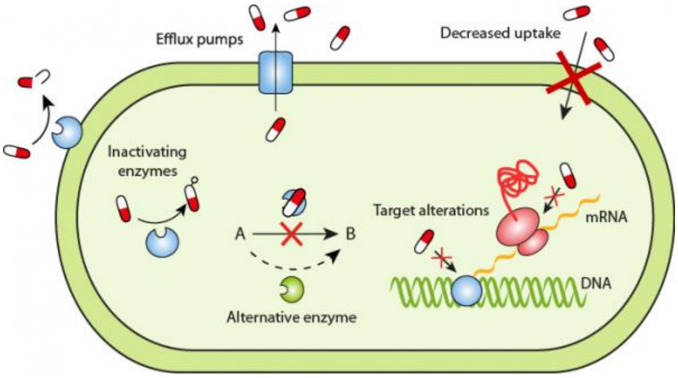


FIGURE 2: **Representation of AR strategies on bacteria.** *Image taken from "https://www.reactgroup.org/toolbox/understand/antibiotic-resistance/resistance-mechanisms-in-bacteria".*

More than 150 antibiotics have been discovered since penicillin was found.[55] However, resistance has emerged to most of them, as represented in FIGURE 3. The rise of these resistant strains has resulted in an increase in morbidity and mortality together with a persistence and remarkable spread of the different resistant populations. Nowadays, AR represents one of the most important worldwide threats to public health.[56]

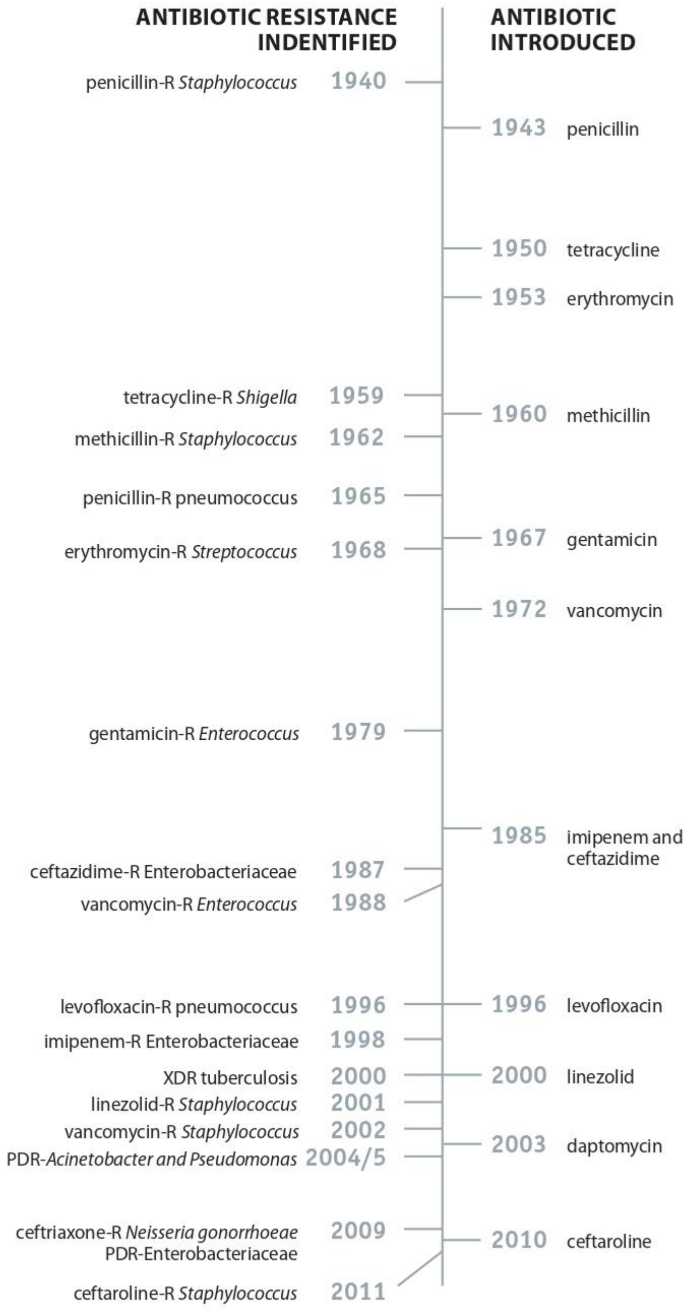

ANTIBIOTIC RESISTANCE INDENTIFIED
ANTIBIOTIC INTRODUCED
penicillin-R *Staphylococcus* 1940
1943 penicillin
1950 tetracycline
1953 erythromycin
tetracycline-R *Shigella* 1959
1960 methicillin
methicillin-R *Staphylococcus* 1962
penicillin-R pneumococcus 1965
1967 gentamicin
erythromycin-R *Streptococcus* 1968
1972 vancomycin
gentamicin-R *Enterococcus* 1979
1985 imipenem and ceftazidime
ceftazidime-R Enterobacteriaceae 1987
vancomycin-R *Enterococcus* 1988
levofloxacin-R pneumococcus 1996
1996 levofloxacin
imipenem-R Enterobacteriaceae 1998
XDR tuberculosis 2000
2000 linezolid
linezolid-R *Staphylococcus* 2001
vancomycin-R *Staphylococcus* 2002
2003 daptomycin
PDR-*Acinetobacter and Pseudomonas* 2004/5
ceftriaxone-R *Neisseria gonorrhoeae* 2009
PDR-Enterobacteriaceae
2010 ceftaroline
ceftaroline-R *Staphylococcus* 2011

FIGURE 3: **Timeline of antibiotic history and when resistance to them emerged.** *Image taken from "Centers for Disease Control and Prevention, from the report Antibiotic resistance threats in the United States, 2013".*

1.3. GENETIC BASIS OF ANTIBIOTIC RESISTANCE

When talking about AR we distinguish two different types. ***i)*** An intrinsic resistance, which concerns bacteria that possess properties that make them naturally resistant to certain drugs. These properties are normally chromosome-encoded and include non-specific efflux pumps, antibiotic inactivating enzymes or some mechanisms that work as permeability barriers.[57] ***ii)*** An acquired resistance, which is generally plasmid-encoded, concerns a naturally susceptible microorganism developing a feature that enables it to not being affected by a drug. Bacteria have an impressive genetic plasticity that enables them to adapt and respond to a wide variety of environmental threats, including antibiotic molecules.[58] As a result, they have been able to survive in the presence of antibiotics using two major genetic strategies to overcome antibiotics: ***i)*** mutations in gene(s) that are normally associated with the mechanism of action the compound has, or ***ii)*** acquisition by horizontal gene transfer of a new and foreign gene that codes for resistance.[41]

The acquired resistance is a more important public health threat as it is normally plasmid-encoded instead of chromosome-encoded, and can therefore spread among bacteria.[59]

1.3.1. Mutational Resistance

Mutational resistance occurs when a given bacterium develops a mutation that affects in any way the activity of the antibiotic. The result of this mutation is the bacterium's survival in presence of the antimicrobial molecule. When this resistant mutant appears, bacteria that have not acquired this mutation will be removed by the compound, therefore, the mutant bacteria population will be the only variant that will remain alive. Usually, these mutations are expensive to cell homeostasis, thus, they are only maintained if required for survival. In general, mutations that result in AR affect the compound action via one of the next mechanisms: ***i)*** modifications of the antimicrobial target (for example, decreasing the affinity for the drug), ***ii)*** mutations that result in a decrease of the drug uptake, ***iii)*** changes in metabolic pathways that modulate regulatory networks, or ***iv)*** activation of a given efflux mechanism that will result in the ejection of the compound. Therefore, AR that arises because of acquired mutations is diverse and complex.[41]

Mutation occurs relatively slowly. The typical bacterial point mutation rate in nature is in the range from 1 in 10 million to 1 billion base substitutions per nucleotide per generation.[60] However, bacteria with about 100-fold higher mutation recurrence can be found in natural and clinical environments.[61] In contrast, rearrangement mutations (insertions, deletions, duplications, inversions) happen at much higher rates, 10^{-5} to 10^{-3} per cell per generation.[62] However, under stress conditions, mutation rates may significantly increase.

1.3.2. Horizontal Gene Transfer

The process of evolution by natural selection continuously generates modifications that help adapt an organism to its environment.[63] When these alterations have a genetic basis, they can be transmitted from parent to offspring, hence, they can be transmitted “vertically”. Nonetheless, it has been demonstrated that genetic information can also be transferred in an inheritance-independent way, a process referred to as horizontal gene transfer (HGT). HGT, also known as lateral gene transfer (LGT) or horizontal DNA transfer (HDT), is the “non-genealogical transmission of genetic material from one organism to another”.[64] By HGT, bacteria are able to respond and adapt to their environments in a more rapid and efficient way than by mutational resistance, as they acquire large DNA sequences from another bacterium in a single transfer.[65,66]

In the microbial world, HGT is a fact of life as it is a leading mode of adaptation, which has an important contribution to genome evolution. Once a new feature that helps survival has been acquired by a bacteria, this may be transferred to its offspring, resulting in them also being more suited for the environment.[67] Most frequently, bacteria adapt to new environmental conditions by the acquisition of genes by HGT rather than by alteration of gene functions through mutations. HGT is the most important factor for genetic modifications in bacteria. Evidence shows that 755 out of 4,288 ORFs (open reading frames) from *Escherichia coli* has been inserted into its genome from the *Salmonella* lineage in at least 234 horizontal transfer events since their divergence.[68]

HGT is responsible for the widespread dispersion of AR genes among bacteria. As already stated before, bacterial resistance to antibiotics is a worldwide health problem.[69] Between 1930 and 1945, the first three classes of antibiotics were being used therapeutically. By 1955, strains of MDR bacteria were reported.[70] and it became clear that the rate in which bacteria were achieving resistance to these antibiotics could not be due to *de novo* mutations alone.[67] In 1960, HGT was proven to be the major mean by which bacteria were exchanging ARGs.[71] Nowadays, virtually all-pathogenic bacteria have been reported to be resistant to multiple antibiotics, for instance, vancomycin-resistant enterococci (VRE), methicillin-resistant *Staphylococcus aureus* (MRSA), multi-drug resistant *Pseudomonas aeruginosa* and *Mycobacterium tuberculosis* being particularly notorious.[69,72,73] Understanding the HGT process in detail is necessary in order to be able to develop new strategies to stop drug-resistant bacteria.

Originally, HGT was thought to occur between closely related bacterial species through three major mechanisms: transformation, transduction and conjugation. These three mechanisms are considered the most important ones up to date. However, we need to be aware that evidence shows that distantly related bacteria can exchange DNA by alternative processes as well.[74]

1.3.2.1. Transformation

Transformation is a process of HGT by which some bacteria take up foreign DNA from the environment, which is generally followed by stable integration of the absorbed DNA in the bacterial genome via homologous recombination.

It was first reported by Frederick Griffith in 1928 using *Streptococcus pneumoniae*.[75] Griffith used two strains of *S. pneumoniae*: III-S (smooth) and II-R (rough). The III-S covers itself with polysaccharide capsule that protects it from the host's immune system while II-R lacks it and thus, it is killed by the host's immune system. Griffith killed III-S bacteria by heat and added their remains to II-R. Neither of these two preparations were able to harm the mice separately, in contrast to the combination of both. Moreover, Griffith was able to purify both live II-R and III-S *S. pneumoniae* from the blood of the dead mice. He concluded that the type II-R had been "transformed" into the lethal III-S by a "transforming factor" that was part of the dead III-S bacteria.

In 1944, Avery, MacLeod and McCarty provided evidence that DNA was the transforming factor that Griffith described.[76] It soon became clear that bacteria could be using this mechanism to evade antibiotics by the exchange of ARGs. In 1951, Hotchkiss induced penicillin and streptomycin resistance in sensitive strains of *S. pneumoniae* by exposing them to DNA from resistant strains.[77] Some years later, it was shown that there could be intra- and inter-species transfer of streptomycin resistance between *Haemophilus influenzae*, *H. parainfluenzae* and *H. suis*.[78,79]

The process of gene transfer by transformation does not require a living donor cell, it only requires persistent DNA in the environment. Actually, virtually all environmental systems contain free DNA, most of it comes from dead cells or broken viral particles, but it has been shown that cells can also actively secrete DNA.[80,81] Moreover, bacteria can take up DNA fragments released by bacteria they themselves have killed.[82] The requisite for bacteria to undergo transformations is the capability to take up free, extracellular genetic material, which is known as being competent. To develop competence, expression of a number of genes encoding DNA uptake and processing systems is activated. Typically, it involves 20-50 proteins that are often composed of components resembling subunits of the type IV pili (T4P) and type II secretion systems (T2SS).[83]

The process of transformation is shown in **FIGURE 4**. DNA fragments are recognized by receptor molecules on the surface. A complex of membrane proteins will transport the DNA across the cell envelope. During this transport, one strand of DNA will be degraded, so a single-stranded DNA (ss-DNA) will be incorporated in the recipient cell. Recombination machinery of the cell may integrate the ss-DNA in the host chromosome. In case the extracellular DNA is a plasmid, it will establish itself and its replication will be independent from the host chromosome.[84]

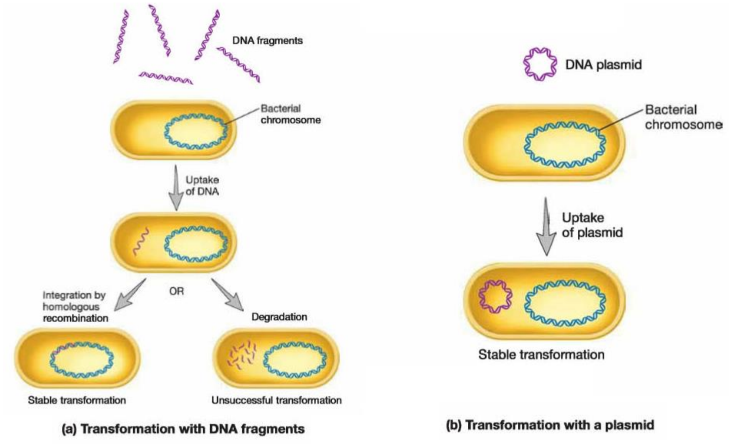


FIGURE 4: **Schematic overview of the transformation process.** The host bacterial cell can take up DNA fragments or plasmids from the environment. DNA fragments can be integrated on the genome by homologous recombination or degraded, whereas plasmids that are transformed, will replicate inside the host cell. *Image taken from "https://istudy.pk/bacterial-transformation/"*

Approximately 60 bacterial species have been documented as being naturally transformable, which represents a very low percentage of all the bacterial species characterized up to date.[85] Nevertheless, competence genes are present on many bacterial species that have not been reported as being naturally competent. This may be due to the fact that natural competence is triggered by certain physiological states that are induced by environmental factors, such as nutrient availability, the presence of peptides or autoinducers, quorum sensing or starvation.[86,87,88] Some examples of competent bacteria are *Escherichia coli*[89], *Bacillus cereus*[90], *Vibrio cholera*[91] or *Bacillus subtilis*[92]. Remarkably, it has been demonstrated that exposure to antibiotics can induce competence in many species, which means that antibiotics do not only lead to resistant strains, but also may lead to incorporation of ARGs.[93,94]

1.3.2.2. Transduction

Transduction is a process of HGT by which DNA is transferred from one bacterium to another one by a virus, known as a bacteriophage. The bacteriophage incorporates bacterial genomic DNA on its capsid and then infects the bacterial cell. Phages are very common and stable on account of their protective coat.[95]

It was discovered by Norton Zinder and Joshua Lederberg on the year 1951 while searching for recombination in the bacterium *Salmonella*.[96] They were using two different strains: *phe$^-$trp$^-$tyr$^-$* and *met$^-$his$^-$*. When only one of these was plated on minimal medium, no wild-type cells were observed, nonetheless, when both of them were plated, wild-type cells were observed. At this point, they thought it was due to the process similar to recombination in *Escherichia coli*.[97]

Nonetheless, they conducted some experiments in which cell contact was prevented by a filter that separated the two arms (commonly known as a U-tube experiment).[97] By changing the pore size of the filter, they realized the agent responsible for recombination was the same size of the virus P22, which is a known phage of *Salmonella*. By further experimentation, they found out that the agent was filterable and had immunity to hydrolytic enzymes and sensitivity to antiserum. This was how Zinder and Lederberg discovered a new type of gene transfer by a bacteriophage instead of recombination in *Salmonella*, which was what they were originally studying.[97]

We now know that there are two types of transduction: ***i)*** generalized and ***ii)*** specialized. In generalized transduction, the bacteriophages can pick up any part of the chromosome. In specialized transduction, the bacteriophages carry only specific portions of the bacterial DNA.[97]

The initial step of infection is the adsorption of phages to host receptors. Phages attach to their host receptor via moieties known as anti-receptors. The specificity of these is flexible, therefore, they can recognize several bacterial surfaces.[98] The phage can enter two cycles of viral replication, the lytic cycle and the lysogenic cycle. Both of them are represented in **FIGURE 5**.

i) In the lytic cycle, the phage binds to the surface of the bacteria and injects its genome into the cytoplasm. Then, phage DNA is copied and phage genes expressed to proteins. In the next step, new phage particles are assembled in the cytoplasm of the bacteria from these newly expressed proteins. The newly formed virus particles are then released through lysis of the cell.

ii) In the lysogenic cycle, the attachment to the receptor and the DNA injection occur the same way as in the lytic cycle. Nevertheless, phage DNA is incorporated in recipient chromosome by recombination. The integrated phage DNA, called a prophage, is not active. However, every time the bacterium is expressing its own chromosomal genes, it will also express the phage genome. Under suitable conditions, the prophage can become active and excise from the bacterial chromosome, triggering the remaining steps of the lytic cycle.

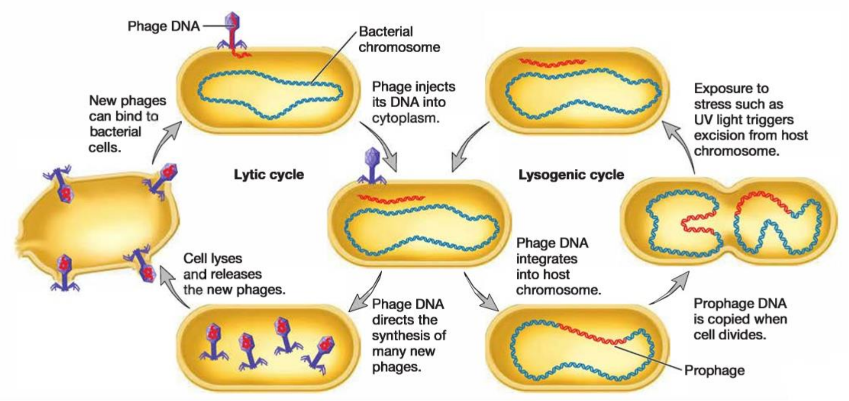

Phage DNA
Bacterial chromosome
New phages can bind to bacterial cells.
Phage injects its DNA into cytoplasm.
Exposure to stress such as UV light triggers excision from host chromosome.
Lytic cycle
Lysogenic cycle
Cell lyses and releases the new phages.
Phage DNA integrates into host chromosome.
Phage DNA directs the synthesis of many new phages.
Prophage DNA is copied when cell divides.
Prophage

FIGURE 5: **Schematic overview of the transduction process.** The first step is the infection of the bacterial cell by the phage, followed by the injection of the phage DNA, both in the lytic and lysogenic cycles. New phage proteins are expressed and assembled and ultimately the cell lyses and the phages are released in the lytic cycle. In the lysogenic cycle, phage DNA is integrated in the bacterial chromosomal and expressed in each cell division. *Image taken from "istudy.pk/bacterial-transduction/"*

The most studied lytic bacteriophages are T4[99] and T7[100], which infect *E. coli*. An example of a bacteriophage known to follow the lysogenic cycle and the lytic cycle is the phage lambda (λ) of *E. coli.*[101]

1.3.2.3. Conjugation

Another process of HGT is conjugation, by which DNA is transferred between bacteria via direct cell-to-cell contact or by a bridge-like connection between two cells. This mechanism has spuriously been compared with eukaryotic sex.

In 1946, Joshua Lederberg and Edward Tatum were trying to elucidate whether bacteria possessed any similar processes to sexual reproduction. They studied two strains of *Escherichia coli* with different nutritional needs; strain *met*$^-$ *bio*$^-$ *thr*$^+$ *leu*$^+$ *thi*$^+$ and strain *met*$^+$ *bio*$^+$ *thr*$^-$ *leu*$^-$ *thi*$^-$. No colonies arose in non-supplemented plates where both strains had been incubated separately. Yet, when mixed together, some of the offspring became wild-type, having reacquired the capability of growing without the need of adding any additional nutrients.[97]

At some point they thought that these two strains were not exchanging the genes but leaking elements that other cells could absorb and therefore, they were able to grow.[102] However, this possibility was rejected by Bernard Davis when he conducted a U-tube experiment where the two arms where both strains tested by Lederberg and Tatum were located in minimal medium. Both strains were separated by a filter whose pore size was too small to let the bacteria or bacteriophages pass but large enough to let medium or any substance pass by. After some incubation time, he did not detect any bacterial growth. This indicated that cell-contact was required for genetic exchange to occur.[103]

Finally, in 1953, William Hayes discovered this type of genetic exchange that needed cell-to-cell contact happened in one direction, in other words, he determined one cell acts as a donor and another cell as a recipient.[97] Hayes noticed by serendipity that a variant of his original donor strain did not produce recombinants on crossing with the recipient strain. The donor strains had lost the ability to transfer genetic material. After further analysis, Hayes concluded that the ability to donate DNA (which he named as fertility) on *E. coli* could be lost and regained easily. He suggested that the ability to donate is a hereditary state based on a fertility factor (F). Strains that have this factor can act as a donor and are therefore called F$^+$, while strains lacking F act as recipients and are labeled as F$^-$.[104]

As mentioned before, bacterial conjugation is the transfer of genetic material between two bacterial cells via direct cell-to-cell contact. This process is schematically shown in **FIGURE 6**. The transfer goes from a donor cell or F$^+$ to a recipient cell or F$^-$. DNA is transferred through a pore that is known as$^-$ the type IV secretion system and connects F$^+$ and F$^-$. The process is characterized by the presence of a conjugative plasmid. Conjugative plasmids contain a number of distinctive elements, including the *oriT* and *TraJ*.

i) *oriT* (origin of transfer): Conjugative plasmids have their own origin of transfer with a nic site, which is a portion of the genome where one of the two strands of the plasmid is cleaved by a relaxase. The ss-DNA is transferred to the recipient cell and the complementary strand is synthesized. The ss-DNA in the donor cell also is converted to double-stranded DNA (ds-DNA). After the conjugative process, both cells can act as F^+.

ii) *tra* (transfer genes): The activation of these genes triggers other plasmid genes that function as proteins that enable direct cell-to-cell contact between two bacteria. These proteins are known as mating pair formation (Mpf) proteins. Some examples are proteins responsible for growing a pilus or proteins that drive the fusion between the outer membranes of two cells.[105]

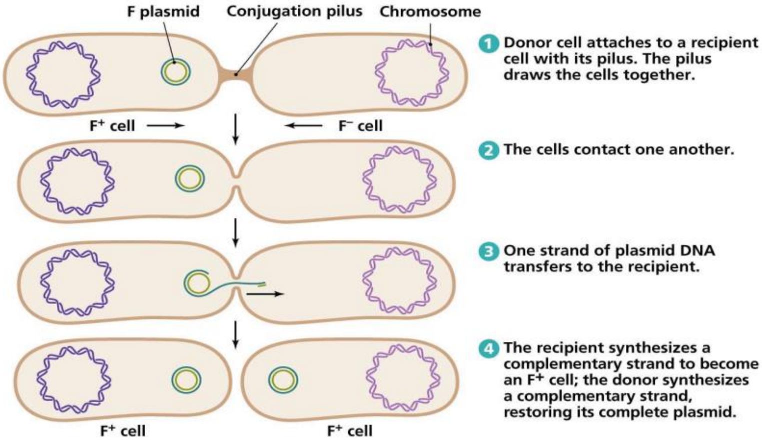

F plasmid
Conjugation pilus
Chromosome
F⁺ cell
F⁻ cell
1 Donor cell attaches to a recipient cell with its pilus. The pilus draws the cells together.
2 The cells contact one another.
3 One strand of plasmid DNA transfers to the recipient.
4 The recipient synthesizes a complementary strand to become an F⁺ cell; the donor synthesizes a complementary strand, restoring its complete plasmid.
F⁺ cell
F⁺ cell

FIGURE 6: **Schematic overview of the conjugation process.** Conjugation is the transfer of DNA from a donor cell to a recipient cell requiring direct contact through a pilus. The conjugative plasmid is cleaved and transferred to the recipient cell as an ss-DNA molecule. The final result is a recipient cell that becomes a donor having its own copy of the plasmid. *Image taken from "study.pk/bacterial-conjugation/".*

Conjugation has the broadest range of host-transfer of genetic material in comparison to transformation and transduction. Conjugative elements have expanded to include not only plasmids, but also conjugative transposons and integrated conjugative elements (ICEs) that are related to bacteriophages.[106]

1.3.2.4. Alternative ways of HGT

Although the afore-mentioned HGT mechanisms have been known for decades, there are some other less-known ways of transfer that do not fit entirely with any of these three canonical processes. Some examples are intercellular connection through nanotubes[107], release of membrane vesicles that contain chromosomal, plasmid and phage DNA that can merge with nearby cells,[108] or temporary cell fusion that leads to chromosomal recombination or plasmid exchange.[109] Another interesting mechanism is mediated by GTAs (gene transfer agents), which are phage-like entities that lack the hallmark capabilities of a phage and they package random pieces of the chromosomal genes of the cell.[110]

1.4. BACILLUS SUBTILIS

Bacillus subtilis is a 4-10 µm sized, rod-shaped Gram + bacterium, capable of growth in the presence of oxygen, which forms a unique type of resting cell called an endospore or spore and therefore allows it to survive in extreme conditions (high temperatures, desiccation, enzymatic and chemical damage or ultraviolet radiation, aerobic and anaerobic conditions, extreme pH, and so on). The cortex, which is a thick peptidoglycan layer, is deposited between an inner and outer membranes.[111] The peptidoglycan layer is in charge of maintaining the core in a dehydrated state that contributes to heat resistance of spores. The outer and inner layers are composed of at least 50 protein species.[112] They work as protective barriers against enzymes or chemicals.[113] Some Bacillus species like *B. anthracis* and *B. cereus* also possess an exosporium surrounding the spore coats. However, in the case of *B. subtilis*, this structure was only observed in some strains isolated from the human grastrointestinal tract.[114] It was first recognized and named in 1872 by the German bacteriologist Ferdinand Cohn. He described the whole life cycle of *B. subtilis* and the formation and germination of endospores by microscope studies.[115] The organism represented what was to become a large and diverse genus of bacteria named *Bacillus*, in the phylum of Firmicutes.

The genus *Bacillus* includes both non-pathogenic and pathogenic species. For instance, *B. anthracis* causes anthrax and *B. cereus* is an agent of food poisoning. In contrast, *B. thurigensis* produces insecticidal endotoxins that are used to control insect pests. *B. amyloliquefaciens* is a source of natural antibiotic protein barnase, α-amylase, Bam HI restriction enzyme and the proteinase subtilisin.

Much of the information we have on the biology, biochemistry and genetics of the Gram + bacteria, has been derived from the study of *B. subtilis.*[116] This is due to its multiple advantages when working with this bacterium. Firstly, it is a non-pathogenic bacterium, generally recognized as safe (GRAS) by the US Food and Drug Administration (FDA). Secondly, all its genomic sequence is available since the year 1997[117] and it is the best characterized species of the genus. Thirdly, *B. subtilis* can spontaneously develop natural competence making it suitable for genetic manipulation.[118] Fourthly, it is closely related to pathogenic bacteria like *B. cereus* and *B. anthracis*[119], and more distantly related to pathogen *Listeria*. Lastly, it can be found in myriad environments, both terrestrial and aquatic, meaning it is ubiquitous and able to grow in diverse settings within the biosphere.

Bacillus species are one of the most widespread bacteria worldwide. As stated before, they produce spores to overcome extreme conditions that enable them to be kept dormant for many years, therefore they can live in a wide range of habitats, mostly soil and sediments. Also, *B. subtilis* strains are gut commensals in animals and humans.[120]

There are several antibiotic-resistance mechanisms in the *Bacillus* species. Some of the features that have been found for *B. subtilis* are macrolide-resistance genes encoded by plasmid *erm(C)* gene,[121] or tetracycline-resistant elements encoded by plasmid *tet(L)* gen[122] and the conjugative transposon Tn*5397.*[123]

Studies have demonstrated that the gut microbiome of humans and animals functions as a pool for ARG.[124] High density of microbes favors HGT, especially conjugation. There are both Gram + and Gram – bacteria in the human gut. A vast part of the Gram + bacteria belong to the phylum Firmicutes.[125] Furthermore, a big fraction of the microbiota in fermented food, which thrives in the gut, also belongs to the same phylum.[126] Because of this, comprehension about how conjugation-mediated transmission of ARG occurs is much needed, especially in Firmicutes, to which it belongs the bacteria *B. subtilis*.

1.5. MOBILE GENETIC ELEMENTS

Pathogens can acquire adaptive phenotypes that allow them to survive in different conditions. These adaptive phenotypes include AR and can be obtained through single nucleotide polymorphisms (SNPs), small insertions and deletions (indels), inversions, duplications, and the movement of mobile genetic elements (MGEs).[127,128]

MGEs are DNA sequences of varying lengths (1 to several hundred kb) that encode enzymes and other proteins that can mediate the movement of DNA within a genome or between bacterial cells. They are now considered key players in the reshuffling of genetic material, and together with mutations and selection, they drive evolution forward. Traditionally, MGEs were classified as bacteriophages, plasmids or transposons. However, this classification became obsolete as new MGEs have been characterized.[129] Here is a brief description of each MGE:

- **Bacteriophages:** Viruses that infect bacterial cells using a protein package known as capsid. Bacteriophages typically carry the genetic information needed for replication of their nucleic acid and synthesis of their protein coats. Normally, the genetic information is DNA, although some RNA phages do exist. When phages infect the host cell, they require its precursors, energy and ribosomes to replicate their nucleic acid and produce the protein coat. Furthermore, even if at low frequency, bacteriophages can accidently package segments of host DNA in their capsid and inject this DNA into a new host, where it can recombine with the cellular chromosome.[130]

- **Plasmids:** Small (average 80kb), extra-chromosomal ds-DNA elements that are capable of semi-autonomous replication through the recruitment of host cell machinery.[131] We will talk in § 1.5.1"Plasmids" about this in deeper detail as this work is focused on proteins of the conjugative plasmid pLS20.

- **Transposons:** DNA fragments that are capable of moving to different DNA sites with varying degrees of site specificity via a recombination event involving an enzyme called transposase. They usually have repetitive DNA sequences at each end to facilitate the excision from the genome. Some excise from the original site and insert into new site (cut and paste), whereas others use replicative mechanisms to create a copy at a new site.[132,133] Transposons are robust forces of genetic change and have had a significant role in the evolution of many genomes. Transposons can also move into phages and plasmids.[134]

- **Integrative and conjugative elements (ICEs) or conjugative transposons (CTns):** Typically mosaic and modular elements, ranging from 20kb to >500kb. They carry genes that encode the machinery necessary for conjugation. They are commonly found integrated in the host chromosome and contain genes required for integration and excision. ICEs are propagated passively during chromosomal replication, segregation and cell division.[106]

1.5.1. Plasmids

Plasmids are key vectors of horizontal gene transfer. They are ds-DNA, self-replicating molecules that are separated from a cell's chromosomal DNA. Although most plasmids are circular, they can be linear too. Plasmids can have varying sizes from 1kb to 300kb. They occur naturally in bacteria, yeast and some higher eukaryotic cells and they live in a parasitic or symbiotic relationship with their host cell. In some cases, plasmids can contribute considerably to the total genetic content of a cell, representing more than a 25%.

Plasmids include essential genes that have replicative and maintenance functions. Apart from those, they can also contain genes that provide some benefits to the host cell to satisfy the plasmid's portion of the symbiotic relationship, which include degradative functions, AR or virulence. For instance, pTOM31c is a 114kb plasmid that contains the TOM pathway. It encodes for toluene *ortho*-monoxygenase (TOM) and catechol 2,3 dioxygenase (C230) genes, as well as for genes required for aerobic, cometabolic mineralization of trichloroethylene (TCE). Moreover, pTOM31c contains a Tn5 transposon carrying the kanamycin resistance marker.[135]

Bacterial plasmids encode for enzymes that inactivate antibiotics. Such drug-resistant plasmids have become a problem in the treatment of several common bacterial pathogens. Due to human overuse and misuse of antibiotics, plasmids containing several drug-resistance genes evolved, making their host cells resistant to a variety of different antibiotics. Some plasmids also contain transfer genes, which encode proteins that can form a macromolecular tube, or pilus, through which a copy of the plasmid can be transferred to other cells. This mentioned transfer results in the spread of drug-resistant plasmids, spreading the number of AR bacteria.[136]

Most plasmids replicate by theta-type or rolling cycle type of replication. Theta-type of replication is named after the Greek character θ as its shape reminds of it when visualized by electron microscopy. A replication initiator (Rep) protein recognizes the origin of the plasmid, binds there, and aids in the melting of the strands. Then, recruitment of the host factors to the origin happens and elongation of DNA synthesis starts. Theta-type replication can be both uni- and bi-directional.[137] Plasmids that replicate via rolling-cycle replication (RCR), contain two types of origins: a double-stranded (dso) and single-strand (sso) origin. Rep protein belongs to the relaxase family of proteins. The Rep protein recognizes dso, introduces a nick in it and binds covalently to this nick site where replication will start in the 3'-OH DNA group. At this step, many other proteins will cooperate like a DNA polymerase III, a helicase and some other ss-DNA binding proteins. The result will be a newly copied ds-plasmid and an ss-molecule. This ss-molecule is then converted into to ds-molecule by host factors at the sso.[138]

Plasmids are used for many different purposes in biotechnology and industry. They were first used as recombinant DNA in the 1970s as a tool to insert genes into bacteria to produce therapeutic proteins.[139] In general, plasmids used for this purpose have been engineered to optimize their use as vectors in DNA cloning. Some of the optimizations include the following features: ***i)*** They are normally much shorter than naturally occurring *Escherichia coli* plasmids, ***ii)*** they include a replication origin, ***iii)*** an antibiotic-resistance gene, ***iv)*** essential nucleotide sequences required for DNA clinging, and **v)** a region in which DNA of interest can be inserted.

One possible classification for plasmids is based on their mode of replication. In this classification, plasmids are divided into two groups: ***i)*** the group that replicates via RCR (smaller than 12kb plasmids from Gram + bacteria), and ***ii)*** the group that replicates according to the theta mechanism (larger plasmids). However, generally speaking, plasmids are classified based on their mobility into conjugative and non-conjugative plasmids. Conjugative plasmids can transfer themselves into recipient cells thanks to self-coding specific enzymes, proteins and a secretion system. In case of non-conjugative plasmids, some of them are mobilizable, in other words, they contain the *oriT* region, and they can be mobilized by a co-inhabitant conjugative plasmid within the same cell. Another type of non-conjugative plasmids can be non-mobilizable ones.[140]

1.5.1.1. Non-mobilizable plasmids

Non-mobilizable plasmids are plasmids that cannot be transferred horizontally to other cells via conjugation. It is estimated that half of all plasmids and, most very large plasmid are non-mobilizable. They need to be transferred by other means of HGT like transformation or transduction. Some examples of non-mobilizable plasmids are *Bacillus subtilis* plasmid pTA1040[141], *Burkholderia phymatum* STM815 plasmid pBPHY01, or *Ralstonia solanacearum* GMI 1000 plasmid pGMI1000MP. Remarkably, these last two are more than 1.9 Mb in size.[142]

1.5.1.2. Mobilizable plasmids

Mobilizable plasmids contain a minimal set of genes that enable them to be transmitted in the presence of a co-inhabitant conjugative plasmid. It was believed that mobilizable plasmids need to have both the *oriT* region and its related MOB gene, which is the gene that encodes for the relaxase.[142,143] However, it was discovered that a relaxase is not necessarily required if the mobilizable plasmid has a simulation of the *oriT* region. Some examples of mobilizable plasmids are pTA1015 and pTA1060 isolated from *Bacillus subtilis* (natto) strains[144,145],and pUB110 from *Staphylococcus aureus*, which can all be transferred by pLS20, a conjugative plasmid found in *B. subtilis*. [146]

1.5.1.3. Conjugative plasmids

Conjugative plasmids are plasmids that encode for all the genes that are required to mediate their transfer from a donor to a host cell. These plasmids are present both in Gram + and Gram – bacteria.[65] The smallest known conjugative plasmid is R338, which is about 34kb in size. Smaller plasmids do not contain conjugative machinery and depend on mobilization or conduction for HGT. Whereas smaller plasmids can be found in hundreds of copies, conjugative plasmids are commonly found in low copy numbers (<10 copies/cell).[147]

1.5. pLS20 PLASMID

Plasmid pLS20 was first identified in Firmicute bacteria *Bacillus subtilis natto* strain IFO3335, which was originally used for fermentation of soybeans to produce “natto”, a well-known dish in South Asia.[148] Therefore, it seems conceivable that pLS20 or relatives play a role in the conjugation-mediated HGT in the gut of humans and animals. In 1977, Tanaka and Koshikawa isolated and characterized the plasmid with several type II restriction enzymes (EcoRI, HindIII, BamHI) so as to be able to start constructing its physical map. Strain 3335 harbors a 65kbp plasmid, pLS20, and a 5.5kbp plasmid, pLS19. Plasmid pLS19 has been reported to be involved in polyglutamate production.[149,150] pLS20 and other plasmids of similar size are present together with 5.4kb-6.0kb plasmids in *B.subtilis natto* strains.[151] pLS20 contains 84 ORFs.[152] pLS20cat, a derivative of pLS20 that carries a chloramphenicol-resistance gene, encodes a protein that suppresses the development of natural competence of its host.[153]

In 1987, Koehler and Thorne showed how the acquisition of pLS20 by *B. subitilis* transcipients rendered this species transfer proficient, providing proof that pLS20 is a conjugative plasmid. Remarkably, it is conjugative both on liquid or solid media. By mixing donor and recipient cells in Luria-Bertani (LB) media for 15 minutes, thousands of transcipients per milliliter are obtained. It can conjugate across several distinct Gram + bacteria that are related to *B. subtilis* such as *B. anthracis*, *B. cereus*, *B. licheniformis*, *B. megaterium*, *B. pumilus* and *B. thuringiensis*.[146] Besides, it has been shown to be able to transfer even chromosomal DNA of a length of up to 113 kb.[154]

pLS20 uses the theta mechanism of replication. By deletion analysis, the fragment that was sufficient for replication was delineated to 1.1kb. This fragment is flanked by two divergently transcribed genes, *orfA* and *orfB*, which are not necessary for replication. *orfA* has homology to the *B. subtilis* chromosomal genes *rapA* (*spoOl*, *gsiA*) and *rapB* (*spoOP*). The minimal replicon contains several regions of dyad symmetry. Some of other features can be ***i)*** DnaA boxes, ***ii)*** an A/T-rich region containing several imperfect direct repeats, or ***iii)*** a replication terminator. The structural organization of the pLS20 minimal replicon is completely different from that of typical rolling circle plasmids from Gram + bacteria and it was suggested that pLS20 belongs to a new class of theta replicons.[155]

In this work, we have studied diverse pLS20 conjugative plasmid-encoded proteins that have different functions in the conjugative process. We have considered pLS20 as good target to study AR for several reasons. *B. subtilis*, the main host of the pLS20 plasmid, belongs to the phylum Firmicutes. As mentioned, the human gut microbiome acts as a pool of ARG with and a big percentage of bacteria in the human gut belong to this phylum. Also, pLS20 has biotechnological interest because of its occurrence in *B. subtilis* natto, which is important in food production. Based on these considerations, we have chosen to focus on the different pLS20 proteins that have specific functions in the conjugative process.

These are the proteins that have been studied in this work:

- **Rco$_{\mathbf{pLS20}}$:** Protein involved in the regulation of expression of the conjugative machinery of pLS20 together with Rap$_{pLS20}$. Rco$_{pLS20}$ binds DNA to inhibit the expression of conjugative genes forming a DNA loop. For DNA loops to occur, the protein needs to form high-order oligomers. We have found and characterized the domain of Rco$_{pLS20}$ involved in tetramerization by X-ray diffraction and we have observed pH-dependent oligomerization states via size exclusion chromatography.

- **Rap$_{\mathbf{pLS20}}$:** Protein involved in the regulation of expression of the conjugative machinery of pLS20 together with Rco$_{pLS20}$, being Phr*$_{pLS20}$its cognate peptide. Rap$_{pLS20}$ acts as a repressor of the repressor of conjugation (Rco$_{pLS20}$), and thus, activates conjugation. When bound to Phr*$_{pLS20}$, conjugation is inhibited. We have analyzed the binding between Rap$_{pLS20}$ and Rco$_{pLS20}$ at different stoichiometries and the effect Phr*$_{pLS20}$ has in complex formation. Furthermore, we have tested the binding between Rap$_{pLS20}$ and different peptides to evaluate the possibility of cross-regulation between different Rap systems. All these experiments were done by the size exclusion chromatography technique.

- **Reg$_{576}$:** Regulatory protein that works in the establishment of the genes once transferred to the recipient cell. Its binding motif in DNA has been determined by size exclusion chromatography and further apo structures of the protein have been solved by X-ray crystallography. One of the structures obtained have been evaluated to have the potential configuration for DNA binding.

- **P34$_{pLS20}$:** Protein in charge of covalent cell adhesion. It is responsible for the initial recognition between bacterial cells to start the transfer of the ARGs. The structure of one of its domains has been elucidated by X-ray diffraction. The domain, named TED, contained a thioester bond and one of these residues was mutated and its structure has also been solved. Furthermore, functionality assays have been performed so as to assess the importance in conjugation of this protein, and specifically of its thioester bond.

By gathering independent information on the processes and their associated proteins, we have shed light into how the conjugative process could be inhibited in order to hinder the propagation of ARGs. Also, by understanding the function of each protein, we are closer to elucidating how pLS20 proteins may work as a whole in conjugation.

2. MATERIALS & METHODS

2.1. PROTEIN CLONING AND EXPRESSION PROTOCOL

All proteins used in this work were previously cloned in Wilfried J.J Meijer's laboratory in the Centro de Biología Molecular "Severo Ochoa" (CSIC, UAM) in the vector pET28b within the lac operon, such that the coding region was placed upstream or downstream of a region coding for six consecutive histidine residues. This 5368bp long plasmid is widely used for bacterial expression and contains a kanamycin resistant cassette.

The expression protocol started with transformation of the clones in BL21-DE3 pLys-S competent *Escherichia coli* expression strain. 1µl of DNA (at about 100ng µl^{-1}) was added to 100µl of competent cells. Subsequently, cells were recovered by an incubation of 2' on ice. 1ml of LB medium was added to the mixture, which was then incubated at 37°C for 1h in agitation (about 200-300rpm). Lastly, 100µl of the mixture were plated on kanamycin-containing LB (LB/kan) agar plates and left overnight (O/N) at 37°C. The next day, two or three colonies were picked, inoculated in 3.5ml of LB/kan and shaken O/N at 200rpm and 37°C. The following day, this pre-culture was added to 1L of kanamycin-containing TB (Terrific Broth) medium and incubated at 37°C and 300rpm until the OD_{600} reached a value of 0.8-0.9. Kanamycin was added at a concentration of 50µg/ml both to LB and TB in all cases. Then, protein expression was induced by adding 1mM IPTG (Isopropyl-β-D-1-thiogalactopyranoside). Expression was allowed to continue at 20°C and 200rpm O/N for approximately 18h. Cells were then harvested by centrifugation at 4000rpm and 4°C for 30'. The bacterial pellet was stored at -80°C.

2.2. CHROMATOGRAPHY TECHNIQUES AND PROTEIN PURIFICATION PROTOCOL

Chromatography is the most widely used biophysical technique that enables the separation, identification and purification of the components of a mixture for a qualitative and quantitative analysis. Chromatography is based on the principle that a mixture, dissolved in a fluid called the mobile phase, is applied onto a surface or into a solid, known as the stationary phase. The components of the mixture travel at different speeds, causing them to separate. This separation is based on various molecular characteristics of the components of the mobile phase and the material of the stationary phase, such as adsorption (liquid-solid), affinity or differences in their molecular weights (M_W).[156,157] Depending on these differences, some components of the mixture remain in the chromatography system for a longer time, while others pass rapidly.

There are various types of chromatography systems, but in this work we used two different types: ***i)*** affinity and ***ii)*** size exclusion chromatography.[158]

Affinity chromatography is based on specific interactions between two molecules, such as interactions between an enzyme and a substrate, a receptor and a ligand, or an antibody and an antigen. These interactions are typically reversible and non-covalent. This technique is known to be the most specific and effective for protein purification. The specific protein that makes a complex with the ligand, which is attached to the solid support (matrix), is retained in the column, whereas proteins that do not bind the ligand (contaminants) leave the column.[159] Frequently the protein of interest is fused to tags that have specificity for a particular matrix. Some of the most used tags are a string of six or ten histidines, known as his-tag or polyhistidine tag, glutathione S-transferase (GST) or maltose binding protein (MBP). For each of these tags there is a specific resin to which the tag binds. These are the nickel-chelating resin, the glutathione resin or the maltose resin, respectively. To recover the protein of interest after all contaminants have been eluted, a component is added to the eluent that competes for the matrix-bound ligands bound to the tagged proteins. In this work, we used histidine tags containing proteins and nickel-chelating resins as matrix. During the purification, a gradient of imidazole is used, in steps, to avoid contaminants. Imidazole competes with his-tagged protein for the binding with the Ni^{2+}-charged resin. A low concentration of imidazole is added to both the binding and wash buffers to interfere with the weak binding of other proteins and to avoid them binding nonspecifically to the column.[160]

Because the purity level of the protein of interest is normally not sufficient for the applications used in this work after the affinity chromatography, we performed a size exclusion chromatography (SEC), also known as gel filtration (GF), afterwards. The basic principle of this method is to separate macromolecules in solution based on their hydrodynamic radius, which in some cases directly correlates with the M_W. In SEC, the stationary phase is constituted of a resin made of polymers that have different chemical composition. These polymers are cross-linked, creating pores whose dimensions depend on the resin. When the elute passes through the column, molecules smaller than the pores diffuse into pores, while bigger molecules simply pass by the pores as they are too large to enter them. The smaller the molecules, the more time they spend inside the pores and therefore, leave the column with proportionally longer retention times. The dimension of the resin pore determines the resolution range of the column, therefore, its pore size is fundamental for a good purification of the target protein.[158]

For the purification of all proteins used in this thesis, the following protocol was used. It consisted on the resuspension of the harvested cells from 1L culture in about 50ml of lysis buffer (500mM NaCl, 20mM TRIS pH8, 5mM Imidazole, one protease inhibitor pill from Merck (cOmplete™, Mini, EDTA-free Protease Inhibitor Cocktail) and sonication (10'' on, 20'' off at an amplitude of 40%) for a total of 5'. This was followed by centrifugation at 18000xg for 30'. The resulting supernatant was filtered using 0.2µm pore size filters. The filtered supernatant was then applied to His-Trap™ High Performance (GE Healthcare) column, previously equilibrated in binding buffer (500mM NaCl, 20mM TRIS pH8, 5mM Imidazole). Once the sample was loaded into the column, elution of the protein of interest was performed by mounting the column on a FLPC machine (Äkta pure). First, a washing step using binding buffer is performed for 10 column volumes. Then, the imidazole concentration of the elution buffer was increased stepwise to elute bound proteins. For this purpose, different percentages of elution buffer (500mM NaCl, 20mM TRIS pH8, 500mM Imidazole) were used: 8% (40mM Imidazole), 20% (100mM Imidazole), 50% (250mM Imidazole) and finally 100% (500mM Imidazole).

The eluted fractions were analyzed using SDS-PAGE 12% or 17.5% (based on the protein size) polyacrylamide separating gel with a 5% stacking gel and run at 200V for 60-90'. The gel was then stained with a solution of Coomasie blue™ (Bio-Rad) for about 30' and destained O/N with a solution containing 50% methanol and 10% glacial acetic acid. The proteins were identified by mass comparison with a standard protein ladder (PageRuler™ Prestained Protein Ladder, 10 to 250 kDa). The purest fractions were selected and concentrated using Amicon® Ultra-15 ml Centrifugal filters (MERCK) with a suitable cutoff regarding each protein's M_W. During and after concentration, the total amount of protein was calculated by measuring the absorption at 280nm using a nanodrop (Nanodrop 2000 by Thermo Scientific). The extinction coefficient (Ɛ) used was calculated from the protein sequence. The presence of nucleic acids in our sample was assessed by measuring the OD relation at 280nm and 260nm (OD_{260}/OD_{280}) of the sample, which is below 0.7 in the absence of DNA.

A second purification step was performed by GF to polish the sample and also get rid of nucleic acids. We worked with ProteoSEC gel filtration columns (Generon) with two different separation ranges: ***i)*** ProteoSEC 3-70HR, for proteins between 3-70kDa and ***ii)*** ProteoSEC 6-600HR, for proteins between 6-600kDa. The buffer used for the gel filtration was 500mM NaCl and 20mM TRIS pH8 for all proteins except for Rap_{pLS20}, for which 250mM NaCl, 20mM TRIS pH8, 10mM $MgCl_2$, 1mM EDTA, 1% glycerol and 1mM β-mercaptoetanol (Rap storage buffer) was used.

2.3. ANALYTICAL SIZE-EXCLUSION CHROMATOGRAPHY

SEC can also be performed for analytical purposes, using columns with small volumes, allowing for quick assessment of the approximate M_W of the components of the sample requiring small volumes. It is based on the same basic principle of preparative SEC and can therefore be used to monitor the quality of a protein preparation, to study complex formation and to identify protein interaction partners and interaction conditions, but the amounts of analyte required are much lower, usually 0.5%-1% of the total column volume at low flow rates. A calibration curve using known M_W standards is needed to estimate the M_W of unknown analytes. For doing so, a linear fit is calculated for the relation between $\log(M_W)$ against V_{el} (elution volume). The equation relating both entities is as follows:

$$y = m \cdot x + b \rightarrow V_{el}\,(MW) = m \cdot \log(MW) + b$$

being m the slope and y the y-intercept

By simple substitution of V_{el} into the equation, the M_W of the molecules that are being assessed can be calculated. In this work, we used a Superdex™ 200 Increase 5/150 GL column (Superdex 200) from GE Healthcare for complex and interactions analysis as well for the assessment of larger proteins' purity level and oligomerization state study, as it has a separation range of 10 to 600kDa. Also, we used a Superdex™ 75 Increase 5/150 GL column (Superdex 75), also from GE Healthcare for determining some smaller proteins' quality level and oligomerization behaviors due to its separation range of 3-70kDa. The values of the slope *m* and intercept *b* in the equation above were determined as:

- Superdex™ 200 Increase 5/150 GL → y = -0,6815x + 5,1906 with an R^2 = 0,93297
- Superdex™ 75 Increase 5/150 GL → y = -1,0928x + 6,3218 with an R^2 = 0,74186

2.4. CRYSTALLIZATION TECHNIQUES

The growth of protein crystals of sufficient quality to allow the determination of the protein structure is the most challenging step of protein crystallization. Crystallization is based on forming solids in which all atoms or molecules are in a highly organized structure known as a crystal. This can be achieved by inducing a sample in a solution to come out of the solution in a specific time window. If the process is too fast, the sample may precipitate. In general terms, the crystallization process is kinetically limited and macromolecules require a supersaturated solution to start to form crystals.[161]

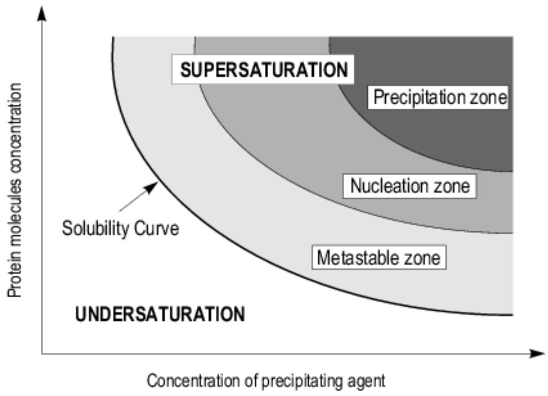


FIGURE 7: **Protein solubility phase diagram.** Schematic representation of a two-dimensional phase-diagram, showing protein concentration versus precipitating agent concentration. Four zones representing different degrees of supersaturation are represented. A supersaturation zone where the protein will precipitate, a moderate supersaturation zone where nucleation occurs, a metastable zone in which crystals can grow but cannot form, and the undersaturation or soluble zone where the protein will never form any crystals. *Image taken from the book "Crystallization of nucleic acids and proteins. A practical approach."*

From the protein solubility phase diagram (FIGURE 7) it is clear that our objective is to promote a supersaturated state, since crystals can only grow in the supersaturated metastable zone. Different techniques are employed to be able to reach this state, which include dialysis, microbatch and vapor diffusion. All crystals obtained in this work were grown using the vapor diffusion technique. For this technique, the protein is in an aqueous buffer and it is allowed to equilibrate with a crystallization buffer in an enclosure. Part of the water of the protein will diffuse to the reservoir. Different crystallization buffers are typically used, corresponding to different crystallization conditions. As a result of the diffusion, the protein drop is dehydrated and a supersaturated state is reached. This technique can be used in two different set ups, hanging drop and sitting drop, which are represented in **FIGURE 8**.

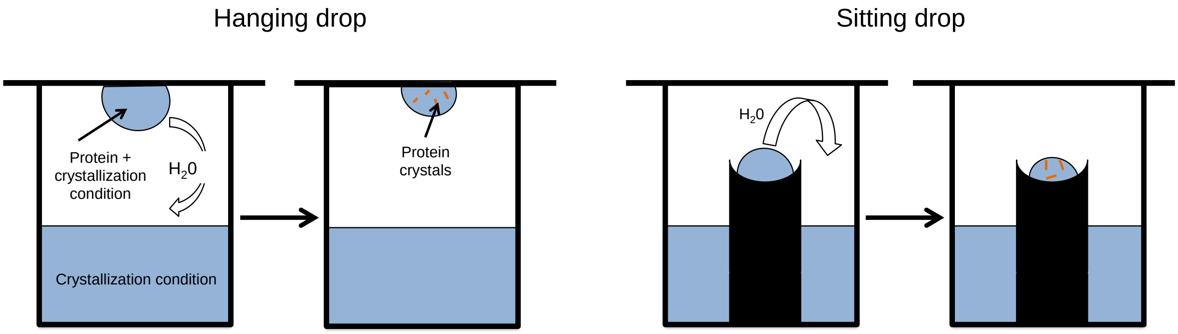

Hanging drop
Sitting drop
Protein + crystallization condition
H_20
Crystallization condition
Protein crystals
H_20

FIGURE 8: **Representation of the two different set ups of vapor-diffusion technique.** In both set ups, a small volume of the protein is mixed with a crystallization solution at a certain radio. This small volume is inverted and suspended above a larger, undiluted solution of the condition in case of hanging drop and placed on a pedestal that is separated from the reservoir in case of the sitting drop.

The initial crystallization experiments are based on a trial and error procedure in which many crystallization buffers, typically hundreds, are used. These buffers are designed to vary as much variables as possible: the type of precipitant (polyethylene glycol, salts,…), concentration of this precipitant, pH, choice of buffer (Hepes, Tris-HCl,…), small additives and so on. In addition, variations in the protein concentration (typically in the range from 10 to 20mg ml^{-1}) and temperature are tried (4°C, 20°C). There are several commercial screening types that encompass a wide variation of these mentioned crystallization buffers. The initial commercial screenings used in this work were the following: PACT premier™, MemGold™, MemGold2™, JCSG+™, Structure Screen 1 and Morpheus® from Molecular Dimensions and Index™ and PEG/Ion Screen™ from Hampton Research. When the crystallization condition is unknown it is common practice to start with many conditions using small volumes (generally 50-200nl). The deposition of drops of such small volumes is made possible through the use of a crystallization robot that is specifically designed and coded for this purpose (Mosquito® crystal from SPT Labtech, Cambridge, UK). Once one or more conditions are found that show evidence of crystal formation, additional crystallization experiments are performed in which the volume of the protein drops is increased to microliter levels, with the aim of obtaining bigger and more ordered crystals. The crystal improvement process is referred to as optimization and it involves sequential changes in the chemical parameters that affect crystallization, such as protein concentration, pH and ionic strength as well as physical parameters like temperature and the kinetics of the equilibrium through changes in the overall methodology. In this thesis, we started performing crystallization trials using the afore-mentioned commercial screening kits with drop volumes of 200nl. Once conditions for good quality crystals were identified, we scaled this condition to drop volumes of 2µl usually introducing sequential changes on the pH and the precipitant percentage.

2.5. STRUCTURE DETERMINATION BY X-RAY DIFFRACTION

When the challenge of growing a crystal of enough quality has been overcome, the structure of the biomolecule can be solved by X-ray crystallography. X-ray diffraction (XRD) relies on the interaction between the electrons in the crystal and photons with X-ray energies, both having dual wave/particle nature, to obtain information about the structure of a crystal and its components.

A crystal is an ordered arrangement of atoms, ions or molecules in a crystalline material. Particles have an intrinsic nature to form symmetric patterns in the three-dimensional space in matter. The smallest group of particles that constitutes the repeating pattern is the unit cell and it contains all the structural and symmetry information. A motive in the unit cell is often present more than once which leads to a higher level of symmetry. The unit cell is defined by three distances (a,b,c) and three angles (α, β, γ).

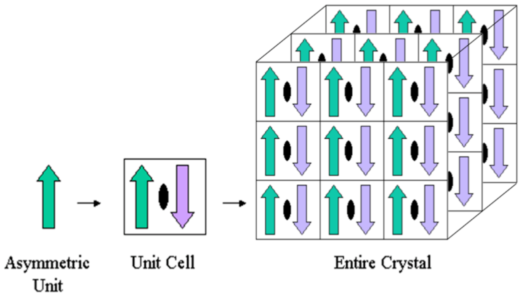


FIGURE 9: **Example of a crystalline structure.** The asymmetric unit is rotated 180 degrees about a two-fold crystallographic symmetry axis (black oval) to produce a second copy. The two arrows together comprise the unit cell. The unit cell is translationally repeated in three directions to make a three-dimensional crystal. *Image taken from "Guide to understanding PDB data"*

When an incident beam of monochromatic X-rays interacts with a crystal, the electrons within the crystal scatter the incident X-rays into multiple specific directions, called reflections. The scattered X-rays undergo constructive and destructive interferences. To be able to generate a three-dimensional representation of the electron density within the crystal two different kind of information are required for each reflection: ***i)*** the phase angle and ***ii)*** the intensity.

The geometry of the diffraction process is described by the Bragg's law:

$$2d\sin\theta = n\lambda$$

In this equation, *d* is the spacing between diffracting planes, θ is the incident angle of the primary beam, *n* is any integer and λ is the wavelength of the beam. Thanks to Bragg's equation, the direction angle of the reflections can be inferred.

When collecting data, the diffracted X-rays hit a detector, which collects the intensities of the reflections. However, diffraction is three-dimensional, while detectors are bi-dimensional. To overcome this and be able to collect all information that comes from X-ray reflections, a highly precise rotation of the crystal is needed. To collect all the reflections that can potentially be generated by the crystal, the crystal is rotated over an oscillation range, while successive diffraction pattern are recorded. Using all of these images, a three-dimensional space, known as the reciprocal space, containing all intensities, is reconstructed. The coordinates of the reflections in the reciprocal space lattice are defined by the Miller indexes (hkl). These indexes represent the set of diffraction planes associated with the reflection and their intensities (I_{hkl}) are determined from the corresponding spot on the detector. The Ewald's sphere model is a mathematical model that represents the diffraction experiments, in which all the feasible reflections are represented as points (the reciprocal lattice) and the geometry of the detector and rotation axis are represented by a sphere. Therefore, the Ewald's construction is a useful tool to visualize the correlation between crystal diffraction and its reciprocal space.

X-ray reflections need to be indexed in order to obtain unit cell's dimensions and the basic symmetry within the crystal, which is known as the space group. From the unit cell and the direction of the diffracted beam, the Miller indexes can be calculated.

To determine the diffracted intensity of each reflection, the recorded intensities of the pixels in the area around the calculated spot position, obtained from the results of the indexing process, are integrated. Subsequently, the intensities are scaled and merged with the objective of correcting errors that may have occurred. With the scaled intensities (I), amplitudes (F) can be calculated using the premise that $I=F^2$. Amplitudes and phases compose the structure factor. Structure factors are waves and as such, can be described by their module |F(h,k,l)| and their phase Φ(hkl). The Fourier Transform (FT) can be used to calculate the three-dimensional distribution of electron density of the unit cell content in real space from the structure factors:

$$\rho(x,y,z) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l} |\, F(h,k,l)| exp[-2\pi i * ((hx + ky + lz) - i\Phi(hkl)]$$

Here, *ρ* is the electron density; *x*,*y* and *z* are the cartesian coordinates in the real space, *V* is the volume of the unit cell, F_{hkl} is the structure factor for a given set of Miller índices, *i* is an imaginary number and *Φ* is the phase. Structure factors are hence, complex numbers. Their real part is related to the intensities measured in the experiment while the imaginary part is related to the phase angle that cannot be directly obtained from the experiment. Since FT requires both intensities (|F(h,k,l)|) and phases (Φ(hkl)) in order to calculate each structure factor, reconstruction of the electron density of a crystal structure requires both values. It is quite straightforward to obtain intensities from diffraction experiments as explained above, but, unfortunately, this is not the case for phase angles. This is known as "phase problem". To be able to overcome this problem, there are four most common fundamentally different structure solution methods:

i) Single and multiple isomorphous replacement (SIR and MIR).

ii) Single and multiple anomalous dispersion (SAD and MAD).

iii) Molecular replacement (MR).

iv) Direct (*ab initio*) methods.

Also, many hybrid structure determination methods can be employed, such as MIRAS. In this worked, we used MR, SAD and *ab initio* phasing program ARCIMBOLDO. These are explained in further detailed.

2.5.1. Single- or Multi-wavelengths anomalous dispersion

Currently, the most common way of solving the phase problem of a novel macromolecular crystal structure is based on the anomalous signal provided by some heavy atoms present in the structure.[162] These experiments can be classified as single- or multi-wavelength anomalous diffraction method. These techniques use one (in case of SAD) or more (in case of MAD) datasets recorded at different energies of the incident X-ray beam from a single crystal that contains suitable anomalous scatterers, which can be introduced by soaking, co-crystallization or biological incorporation of modified amino acids.

Anomalous scattering occurs when well-ordered atoms in the crystal absorb a part of the X-ray energy and thereby change the phase of the scattered wave. The effect is maximal when the X-ray energy is just below or equal to the energy of electrons of the irradiated atoms (absorbing edge). However, anomalous scattering occurs basically always. Even the lightest elements like carbon or nitrogen in a structure absorb the X-rays, yet this contribution is too small to attribute to any individual atom. If a heavy atom is present in the crystal, this anomalous scattering modifies the atomic form factor or atomic scattering factor f(S) sufficiently for the intensity differences to be measured. The atomic scattering factor is a measure of the scattering amplitude of a wave of the element. Protein phases can be estimated from the wavelength-dependent quantitative differences in the anomalous scattering contribution of these heavy atoms in the crystal.

In case of MAD, a single crystal is diffracted multiple times at different beam energies. Most often, three energy points are selected to maximize the specific anomalous scattering, which correspond to the energy of absorption edge inflection point, the energy of the absorption edge peak (maximal absorption) and an energy value aside absorption edge (remote energy).

The major disadvantage of this technique is the effect of radiation damage as the crystal is diffracted a number of times. Therefore, two-wavelength MAD experiments are often performed, as they may offer the best compromise between phase quality and minimizing the effects of radiation damage to the sample. Using SAD, it is possible to obtain limited phase information from a single dataset measured at the energy of the edge of the anomalous scatterer. Nowadays, selenium-SAD is one of the most used experimental phasing techniques due to relative ease of incorporation of selenomethionine into recombinant proteins. One of the main advantages of SAD is the minimization of time that the crystal is exposed to the beam, which considerably reduces potential radiation damage when collecting data.

2.5.2. Molecular replacement

Molecular replacement exploits the structural similarity between the protein of study from which the diffraction data is derived and homologous proteins for which protein structures already exist.[163] MR enables the solution of the crystallographic phase problem by providing initial estimations of the phases of the new structure from a previously solved one, and does not require special data collection experiments, in contrast to SIR/MIR and SAD/MAD. The use of MR has become increasingly more frequent as the database of known structure grows. Approximately 70% of deposited macromolecular structures have been solved by MR. Its main advantages are that it is highly automated, fast and cheap.

The principle of MR is quite simple. Having measured a set of diffraction intensities of our protein, they are compared to the calculated diffraction intensities of all possible orientations and positions of the model in the asymmetric unit of the measured data. By doing so, a predicted diffraction that best matches the observed diffraction is found. The phases of the unsolved crystal are "borrowed" from the phases calculated from the model protein as if it were the model protein that had crystallized in the unknown crystal. This way an initial map is calculated, based on the borrowed phases and the measured amplitudes. The success of the method depends on the interpretability of the corresponding electron density map. If the map can be interpreted, the model can be further improved and refined as described below.

In order to define an orientation and position of the homologous model in the unit cell of the unknown structure six parameters are required: three rotational angles (α, β, γ) and three translation axes (t_x, t_y, t_z). If the asymmetric unit is composed of N molecules, then 6N parameters are needed to determine the solution. Since this search would take a long time, most programs split it in two parts: First, a rotational search is performed and once the best solutions are obtained, a translations search is done.[164]

2.5.3. Direct (*ab initio*) methods

Direct (*ab initio*) methods refer to processes that intend to derive a crystal structure model from the observed structure factors. They function with as little structural information as possible, normally considering the chemical content of the unit cell alone. This is now routinely used for small molecules at high resolutions, since a limited number of structural conformations fit with the experimental data.[165] However, for macromolecules, theses methods can only be used for favorable cases, although recent advances have been through the combined use of small secondary structure elements and a robust MR approach as implemented in *ARCIMBOLDO*, an *ab initio* crystal structure solution program[166,167].

ARCIMBOLDO[166,167] constitutes a phasing method for macromolecular crystallographic X-ray diffraction data at resolutions of 2Å or better. For doing so, it combines the search for model fragments using the MR program PHASER[168] and density modification and main chain auto-tracing using the program SHELXE.[169] It was named after the Italian painter Guiseppe Arcimboldo, who is known for creating portrait heads made of objects such as fruits, vegetables, flowers or books.

Following this analogy, ARCIMBOLDO constructs unknown structures by association of small secondary structure elements. Phasing a structure form partial information provided by such a small percentage of the total model (approximately 10% of the main chain atoms) is challenging and evaluation of alternative hypothesis under statistical constraints is necessary. A multi-solution approach is therefore applied, which evaluates many possible solutions because of the difficulty to recognize correct solutions at early stages.

For complex structures (over 400 residues), location of secondary structure elements is not recommendable unless very long helices are present and high resolution data is available. In these cases, a powerful parallel computing grid is required to distribute calculations over multiple processors. For structures with up to 200 amino acids in the asymmetric unit, a single workstation is enough.

2.6. PHASE OPTIMIZATION AND STRUCTURE VALIDATION

The initial phases obtained from the structure solution methods enumerated above are normally not very accurate. Hence, refinement methods are performed to improve the phases and the interpretation of the electron density map, and therefore the final structure. To do so, a good fit between the experimental data (F_{obs}) and the calculated model (F_{calc}) must be achieved using iterative cycles of manual building and automatic refinement of the model against experimental data.

Work R-factor (R_{work}) is a parameter that measures the agreement between the crystallographic model and the experimental X-ray diffraction data. It can be used to determine the fit between the model and the experimental data and through refinement of the model and the phases, be improved. The R_{work} can get to a local minimum giving the false impression of having a good model, a phenomenon called model bias, which is usually due to over-refinement of the model. Thus, the free R-factor (R_{free}) is used to assess the degree of model bias and thereby validate the model. R_{free} is computed similarly to R_{work} but based on a small set of data that are not included in the refinement, therefore it gives an independent measurement of the fit of the observed and calculated structure factors. R_{free} is always higher than R_{work} as the model is not adjusted to fit these reflections. Since a correct model should predict all data with uniform accuracy, the calculated structure factors should have a similar value to the experimental structure factors for both the working and free set. The divergence between R_{work} and R_{free} are therefore a measure of over-fitting and model bias.

Refinement can be done following two different approaches: ***i)*** maximum likelihood and ***ii)*** simulated annealing, although they can be combined. Both methods use restraints that limit the parameters space of an atomic model to ideal values, obtained from accumulated structural knowledge. These restraints are normally applied to bond distances, angles and torsions and temperature factors (B-factor), to name a few. In maximum likelihood the model and phases are adjusted to minimize the R-factor, whereas in simulated annealing some randomness is first added to the atom positions of the structure to then refine it. This randomness reduces the probability of having a model with a wrong local minimum value.

The result of structure refinement, the refined model, should be thoroughly validated for its stereochemistry, interatomic distances, bond distances, main chain torsion angles, Ramachandran outliers, correct rotamers conformation and position of the water molecules.

2.7. BIOINFORMATICS TOOLS

Bioinformatics can be defined as "the application of computational tools to organize, analyze, understand, visualize and store information associated with biological macromolecules.[170] Important biological questions can be addressed by bioinformatics that include the comprehension of genotype-phenotype connection in human diseases, understanding the structure to function relationship for proteins and learning about biological networks.

In this thesis, we have conducted bioinformatics homology searches based on sequence of a protein to find potential homologs and structure alignments to learn about a newly structurally characterized protein.

2.7.1. Sequence alignments

Pairwise sequence alignment is used to identify regions of similarity between two biological sequences (protein or DNA). Similarities might be due to functional, structural or evolutionary relationship between those sequences. Multiple sequence alignments (MSA) are the alignment of three or more biological sequences. Based on the results, homology and evolutionary relationships between the sequences can be inferred. All sequence alignments performed in this work were done by Clustal Omega from EMBL-EBI.[171] Results were viewed by Jalview.[172]

Furthermore, BLAST (Basic Local Alignment Search Tool) from NCBI was used. BLAST is an algorithm for comparing primary biological sequence information, protein sequences or DNA/RNA sequences. It allows the search of a subject protein or nucleotide sequence, known as a query, against a database of sequences. It identifies the matches above a certain threshold and allows the identification of members of gene families or the establishment of functional and evolutionary relationships.[173] Different types of BLASTs are available. In this thesis we worked with protein BLAST, running a blastp (protein-protein blast) algorithm.

2.7.2. Structural alignments

In order to compare the structures of proteins obtained in 3D with all deposited structures in the PDB (Protein Data Bank) two bioinformatics tools were used: PDBeFold[174] from EMBL-EBI and Dali server[175] from the University of Helsinki, Holm group and Biocenter Finland. By these, we are able to identify structures similar to that of the reference protein. It examines the protein structure for similarity using the full PDB or SCOP (Structural Classification Of Proteins) as a reference repository. It is a powerful structure alignment service that can perform pairwise and multiple three-dimensional alignments. This tool is useful to identify similar proteins in the PDB and, thus, to infer a possible function to it, which may not have been detectable by sequence comparison. Unlike sequence alignment it is not residue-based, but based on identification of residues occupying “equivalent” geometrical positions. There are several options in which the results can be sorted: ***i)*** Q score (Cα-alignment), ***ii)*** RMSD, ***iii)*** Z score (which is based on Gaussian Statistics), ***iv)*** P score (which depends on RMSD, number of aligned residues, number of gaps, number of matched secondary structure elements and the SSE match score), and ***v)*** sequence identity

CHAPTER 1: Rco$_{pLS20}$, Inhibitory protein of pLS20 conjugation

C 1.1. OBJECTIVES

$\text{Rco}_{\text{pLS20}}$ is a key player in the regulation of genetic switch as it drives the inhibition of conjugative genes expression in pLS20. Little is known about how $\text{Rco}_{\text{pLS20}}$ exerts its inhibitory effect apart from binding as a tetramer and forming a DNA loop in the promoter region. By obtaining information about $\text{Rco}_{\text{pLS20}}$ structure, we would gather information of extreme importance to understand how it binds to its promoter region and to $\text{Rap}_{\text{pLS20}}$, which is the activator of conjugation. By being able to boost the binding between $\text{Rco}_{\text{pLS20}}$ and its promoter region or hinder that of $\text{Rco}_{\text{pLS20}}$ with $\text{Rap}_{\text{pLS20}}$, we would be able to inhibit conjugation and thus, the spread of ARGs. We already have structural data of $\text{Rap}_{\text{pLS20}}$ but not of $\text{Rco}_{\text{pLS20}}$. Structural data would aid in finding ways to increase or disrupt interactions that $\text{Rco}_{\text{pLS20}}$ establishes.

These are the concrete objectives of this chapter:

- Obtaining the atomic structure of apo $\text{Rco}_{\text{pLS20}}$ and bound to DNA.
- Comparison of the obtained structure with inhibitor proteins of other conjugative plasmids.

C 1.2. INTRODUCTION

As already stated in the general introduction, conjugation is the major determinant in the spread of genes among bacteria, endowing recipient cells with new catabolic pathways, antibiotic resistance, or virulence.[83] These genes can be carried in plasmids, but they can also be found as MGEs that are integrated in the bacterial chromosome. These latter forms are generally ICEs. Conjugative ransfer of ss-DNAs occurs via type IV secretion system[176] and it needs to be strictly regulated.

It is very common that transcriptional regulators are produced in limited amounts per cell. This results in an increased probability of transcriptional fluctuations between individual cells within a population. Moreover, intrinsic stochasticity also contributes to these mentioned fluctuations.[177] Some differentiation processes like development of natural genetic competence, sporulation or swimming are induced by stochastic variability in levels of transcriptional regulators. It has been demonstrated that bacteria show a high amount of phenotypic diversity under the same conditions.[178] Stochastic fluctuations can result in the presence of two different subpopulations, which is commonly known as bistability.[179] This switching mechanism may have emerged from the need of bacteria to overcome disadvantageous conditions. Hereby, some cells are always disposed to deal with conditions that may be harmful to them.

Nevertheless, the fact of having cells prepared for future conditions is not so beneficial in some other cases and the processes should be inhibited when they are not needed. Conjugation is one of these processes as it is not beneficial for the fitness of the species to have conjugation genes activated by default when there is no need for this function the most of the time. This is in line with the fact that efficiency of pLS20 transfer in growth conditions is at least six orders of magnitude lower than those observed during optimal conjugation conditions, revealing conjugation genes are repressed under growth conditions.[180]

Conjugation is a very complex and energy consuming HGT process as it involves the generation and transfer of ssDNA, synthesis and assembly of a type IV secretion system and finally, the establishment of proper contacts with the recipient cell. Thus, the expression of genes taking part in the conjugative process needs to be tightly regulated.

The regulatory circuitry of pLS20 conjugation genes has been studied.[181] The conjugation genes of pLS20 are located in a single operon that contains over 40 genes (flanking genes *28* to *74*). Gene *27* codes $\text{Rco}_{\text{pLS20}}$, the master regulator of the process. $\text{Rap}_{\text{pLS20}}$, coded by gene *25*, is the anti-repressor that is required for inhibiting $\text{Rco}_{\text{pLS20}}$ and thus, activating conjugation. Gene *26* codes for the short peptide $\text{Phr*}_{\text{pLS20}}$, which controls the activity of $\text{Rap}_{\text{pLS20}}$. $\text{Rco}_{\text{pLS20}}$ was predicted to belong to the Xre-family of transcriptional regulators and it is predicted to contain a Helix-Turn-Helix (HTH) binding motif in its N-terminal part.[181] Conjugation genes are under the control of a genetic switch composed of three intertwined layers.

The first level of regulation concerns the positioning of P_c, which is the main conjugation promoter and the divergently orientated P_r, driving the expression of rco_{pLS20} gene. The presence of divergently orientated promoters is a prevalent manner of gene organization in bacteria. P_c is a strong promoter that partially overlaps with the weaker P_r promoter. Since RNA polymerase can only bind one of the two promoters at a time, it only drives the transcription of a gene controlled by a single promoter. Overlapping promoters have different activities in the presence of the regulatory protein, which may drive a transcriptional switch.[182] In case of P_c/P_r, in the absence of $\text{Rco}_{\text{pLS20}}$, P_c is several hundred times stronger than P_r, while in the presence of $\text{Rco}_{\text{pLS20}}$, P_c is repressed and P_r activated.

$\text{Rco}_{\text{pLS20}}$ binds to two operator sites; O_I and O_{II}. O_I is located more than 85bp downstream of P_c, while O_{II} is located near promoter P_c/P_r. Therefore, the intergenic rco_{pLS20} gene region contains two operators that are separated by 75bp. Each operator site has repeats of a motif 5'-CAGTGAAA-3, which are included in the binding site of $\text{Rco}_{\text{pLS20}}$. Motifs in O_I are placed in the lower strand, whereas motifs in O_{II} are placed in the upper strand, with the exception of one.[180]

The second level of regulation involves the dual effect $\text{Rco}_{\text{pLS20}}$ has on P_c/P_r. As mentioned above, $\text{Rco}_{\text{pLS20}}$ inhibits the strong promoter P_c and triggers the activation of the weak promoter P_r, thereby inducing its own expression, causing a feedback loop that keeps the conjugation in an "OFF" state. Furthermore, repression of the P_c promoter is obtained by a ten-fold lower induction compared to the activation of P_r promoter.[180] Remarkably, Rco_{pLS20} was shown to inhibit its own transcription at high concentrations. By this means, Rco_{pLS20} is maintained in a low concentration range so as to be able to properly react to anti-repressor Rap_{pLS20} and activate conjugation. Once Rap_{pLS20} induces the conjugative process to an "ON" state, it relieves the repression of the P_c promoter and extinguishes the autoestimulation of the P_r promoter, which is translated into the prevention of further synthesis of Rco_{pLS20} and thus, triggers the maintenance of the "ON" state.

The third level of regulation of the genetic switch that activates conjugation is related to the DNA loop that it is formed when Rco_{pLS20} binds to both operators O_I and O_{II}. DNA loops occur when a protein or a complex of protein bind to two distant sites of DNA, looping out the interceding DNA. DNA looping mediated by transcriptional regulators is a common form of regulation in prokaryotes, such as the regulation of the *lac* operon in *E. coli*.

There are several factors that are required for the formation of the DNA looping. Importantly, these requirements are fulfilled in case of Rco_{pLS20} and the DNA region between P_c/P_r. Firstly, Rco_{pLS20} is predicted to contain a HTH DNA binding motif in its N-terminal part[181] and specifically and cooperatively binds to operators O_I and O_{II}. Secondly, operator O_I is needed for proper regulation of both P_c/P_r and it is located more than 85bp upstream from them. Moreover, dephasing the positions of these promoters by 5bp inhibits the expression of the conjugation genes. Lastly, Rco_{pLS20} has been shown to form tetramers in solution, generating a unit containing various DNA binding motifs and binding cooperatively to DNA.[180]

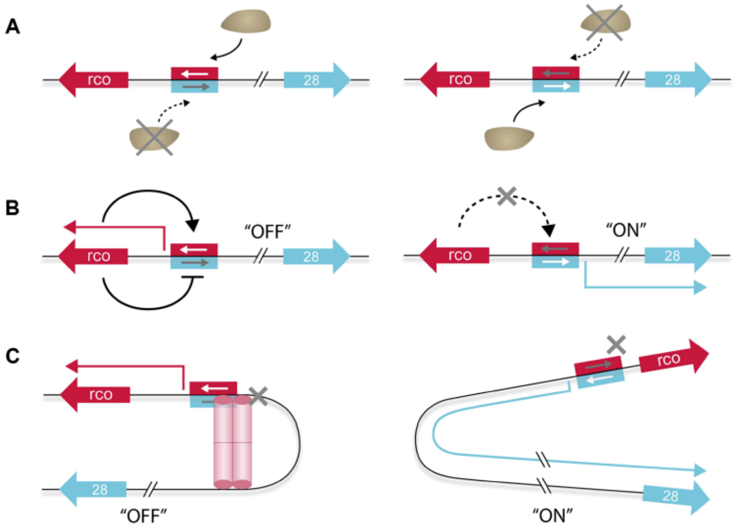


FIGURE 10: **Model of the different regulation layers that contribute to the genetic switch that controls pLS20. A)** RNA polymerase acting as a switch due to its inability to bind both promoters at the same time. **B)** Rco_{pLS20} activates P_r and represses its own promoter P_c, resulting in a positive feedback that maintains conjugation in an "OFF" state. When P_c promoter's repression is relieved and P_r inhibited, conjugation turns to an "ON" state. **C)** High concentration of Rco_{pLS20} is located in the P_c/P_r area due to DNA looping, keeping the P_c promoter repressed and conjugation in an "OFF" state. This repression can be quickly reversed to the activation of expression of conjugation genes if needed. *Image taken from "doi: 10.1371/journal.pgen.1004733"*[180]

The first time that the possibility of a transcription factor could simultaneously bind to two distant sites arose in 1977 by a work from Kania and Muller-Hill.[183] The first experimental proof came in 1984 and it became clear how DNA looping could have an essential role in transcriptional regulation.[184] This demonstration concerned the *ara* operon from *E. coli*, and since then, DNA looping has been shown to transcriptionally regulate other operons.[185] Nevertheless, studies demonstrating DNA looping in plasmids are scarce. One of the best characterized is the regulation of initiation of DNA replication at the beta origin of R6K plasmid from *E.coli*.[186] Regulatory mechanisms that involve DNA looping have also been found in eukaryotes. Usually, require long-range interactions and form higher-order chromatin structures.[187]

DNA loops can be classified into two different groups based on the spacer length: ***i)*** the short or energetic loops or ***ii)*** long and entropic loops.[180] In case of pLS20, O_I and O_{II} are separated by 75bp, thus it belongs to the first group. In such systems, the energy required for formation of the DNA loop is considerable due to the fact that DNA is naturally stiff and hard to torsion. Hence, this process results disadvantageous and for it to occur, intrinsic static bent or binding of an additional protein that provokes the bending are needed. In pLS20, the presence of a static bent in the spacer region between regions O_I and O_{II} has been demonstrated.[180]

It has been experimentally shown that DNA looping provides several advantages. The major benefit comes from the high local concentration of the regulatory protein at the right place.[188] Similar levels of repression could be obtained by a higher concentration of the regulatory protein, yet this will increase the probability of nonspecific binding to unrelated sites. Through DNA looping, a specific binding with a lower concentration is accomplished.[189] Furthermore, it is remarkable how many transcriptional regulators, including Rco_{pLS20}, are produced in a limited manner as mentioned above.

Another advantage of DNA looping is the attenuation of fluctuations. Gene regulation has a stochastic component that may have an important impact in biological processes.[190] Noise comes from the limited amount of molecules involved in the regulation of individual cells. DNA looping has been suggested to be one of the mechanisms developed to attenuate these fluctuations.[191] By DNA looping a fast switch between active and inactive state is enabled as the repressor can be easily captured by the operator. On the contrary, in systems with slow switches, mRNA is produced during long periods of time or not produced at all, which causes fluctuations in the expression levels. Furthermore, formation of the DNA loop can prevent the access of other regulators that may obstruct the regulation process.

Rco_{pLS20}-mediated repression is relieved through direct interaction with Rap_{pLS20}, which is blocked in the presence of Phr^*_{pLS20}. Thus, activation of conjugation is regulated by Phr^*_{pLS20} concentrations. [192,180,181] The regulatory circuitry is represented in **FIGURE 11**.

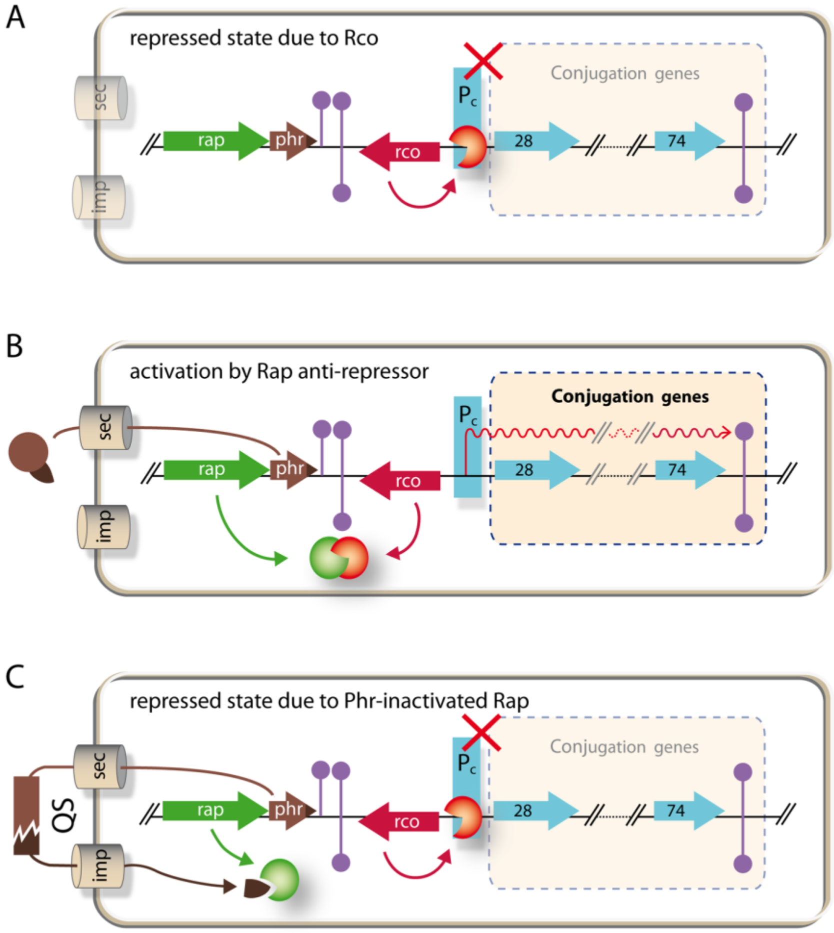


FIGURE 11: **Schematic representation of regulatory circuitry of *B. subtilis* pLS20 conjugation genes.** A) Rco$_{pLS20}$ is bound to DNA repressing conjugation. Gene *rco$_{pLS20}$* (red arrow) is divergently transcribed from the putative conjugation operon enclosing genes 28 to 74 (blue arrows). Rco$_{pLS20}$ inhibits expression of conjugation genes by binding to promoter P$_c$, which is located upstream gene 28. B) Activation of conjugation by anti-repressor Rap$_{pLS20}$. Gene *rap$_{pLS20}$* (green arrow) encodes anti-repressor Rap$_{pLS20}$, which binds to Rco$_{pLS20}$, and thus, inhibits the binding between Rco$_{pLS20}$ and P$_c$ promoter, activating conjugation. C) Inhibition of conjugation by Phr*$_{pLS20}$ peptide binding to Rap$_{pLS20}$. Gene *phr$_{pLS20}$* (brown arrow) encodes a pre–protein of 44 residues. This pre-protein matures outside the cell (as represented by the brown interrupted rectangle) by secretion-maturation-import system (grey cylinders) into a pentapeptide that inhibits Rap$_{pLS20}$ by binding to it. *Image taken from "doi: 10.1371/journal.pgen.1004733"*[180]

The concentration of the signaling peptide is high when donor cells are predominantly surrounded by donor cells, whereas the concentration is low when donor cells are surrounded by recipient cells. As a result, conjugation is active when recipient cells are present. Therefore, Phr*$_{pLS20}$ has an essential role to return conjugation to the default "OFF" state.[193] Moreover, preventing expression of conjugation genes when recipient cells are not present, may also be useful for other aims. For example, pLS20cat replicates via the theta mode of replication under normal circumstances[155], but during conjugation it changes to rolling circle mode so as to generate the ss-DNA that is transferred into the recipient cell. These two types of replication are likely not compatible and by an strict regulation of the conjugation genes, among which are those that initiate rolling circle replication[138] and so, contribute to following the mode of replication depending on the circumstances.

Although Rap$_{pLS20}$-Phr*$_{pLS20}$-mediated regulation is similar to already known mechanisms involved in sporulation or competence development,[194] the knowledge on the regulation of the genetic switch responsible for activating conjugation remains limited and scarce. Even if some regulatory mechanisms have been identified, it is likely that the pLS20 conjugation genes are regulated by further mechanisms as well. Furthermore, studies about the regulatory roles of Rco$_{pLS20}$ and Rap$_{pLS20}$-Phr*$_{pLS20}$ separately have been carried out,[180,195] little has been researched about the interactions between these two proteins. The Rap$_{pLS20}$ structure has already been obtained with and without the Phr*$_{pLS20}$ peptide. However, the atomic structure of Rco$_{pLS20}$ remains unknown, which could shed some light into the feasible interactions that may occur between them.

In this chapter, we describe the structure of a part of the Rco$_{pLS20}$ regulatory protein, more concretely of its oligomerization domain (OD), which we found to form tetramers. This domain was not identified from the primary sequences. The obtained structure showed that it is a structural analog of human oncogene p53, and constitutes the first example of such a structure in bacteria. The OD of Rco$_{pLS20}$ is composed of a dimer of dimers that is stabilized by salt bridges. We also studied the oligomerization state of Rco$_{pLS20}$ was analyzed at different pHs.

C 1.3. MATERIALS AND METHODS

C 1.3.1. Crystallization of Rco$_{pLS20}$Tet

Rco$_{pLS20}$ was purified as described in §2.2 "Cromatography techniques and Protein purification protocol". A second purification based on its particle size (M$_W$: 20kDa) was done using a ProteoSEC 6-600HR column (Generon) using 250mM, 20mM TRIS pH8 as a buffer. Afterwards, it was concentrated to 17mg ml^{-1} over an Amicon® Ultra 15mL centrifugal filter with a cutoff of 10kDa (Merck). Then, it was gently mixed with previously annealed double-stranded oligonucleotide with forward sequence 5'-GTCAGTGAAAAA-3' at a 2:1 (protein:DNA) stoichiometry. Crystals giving the highest resolution were obtained by the sitting drop vapor-diffusion method at 18°C, by equilibration of drops of 100nl of protein + 100nl of crystallization buffer (0.1M Hepes (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) pH 7.5, %28 Polyethileneglycole (PEG) 600) against 100 µl of crystallization buffer in the reservoir. Crystals were harvested after three months of incubation and they were fished and cryo-cooled by direct transfer from the crystallization drop into liquid nitrogen for X-ray collection.

C 1.3.2. Diffraction and data processing of Rco$_{pLS20}$Tet

Data collection was performed at ALBA Synchrotron Light Source on the BL13-Xaloc beamline.[196] Data was processed the AutoPROC toolbox (Global Phasing Ltd.[197]), using anisotropic resolution cutoffs.[198] The structure was determined *de novo* using ARCIMBOLDO[166], followed by automated model building in PHENIX 1.18.2[199]. Refinement was also done in PHENIX 1.18.2 with a cutoff high-resolution of 2Å. Figures were prepared using PyMOL (The PyMOL Molecular Graphics System version 2.3 Schrödinger, LLC). Superpositions were done by the cealign function in PyMOL.[200] F_o-F_c maps were prepared with CCP4i.[201]

C 1.3.3. Analytical SEC on Rco$_{pLS20}$ at different pHs

To study the effect of the pH on Rco$_{pLS20}$ oligomerization state, 25µg (1.2nmol or 49.4µM) of Rco$_{pLS20}$ were injected on a Superdex 200 column. Equilibration of the column was performed at different buffers. For pH5, 500mM NaCl, 20mM citrate buffer pH5 was used. For pH8, column was equilibrated in 500mM NaCl, 20mM TRIS-HCl pH8. For pH10, 500mM NaCl, 20mM Gly-NaOH pH10 was used. Elution was done at a flow rate of 0.2ml min^{-1}. 25µl were injected in all cases. Detection was performed using UV-Vis (ultraviolet-visible) spectroscopy at 280nm and 260nm. To estimate the M_W, a calibration of the column was performed as described in §2.3 "Analytical Size Exclusion Chromatography". The derived relation between the V_{el} and M_W was $V_{el} = -0.6815 \cdot \log(M_W) + 5.1906$, with an $R^2 = 0.933$.

C 1.3.4. Structure and sequence alignment of Rco$_{pLS20}$Tet

Different bioinformatics tools were used to pursue the analysis of the solved structure. On the one hand, PDBeFold from EMBL-EBI was used to perform a structure comparison. The comparison was performed using as a target the whole PDB archive, setting the lowest acceptable match to 70%. Furthermore, protein BLASTs were also carried out using both Rco$_{pLS20}$FL sequence and the structurally characterized part's sequence. Searches were done using non-redundant protein sequences as the database and blastp (protein-protein BLAST) as the algorithm. General searches as well as searches using specific organisms (Archaea, Fungi/metazoan, *H. sapiens*) were performed.

C 1.4. RESULTS

C 1.4.1. Crystallization of $\mathrm{Rco}_{\mathrm{pLS20}}$

To shed light on the structural basis of $\mathrm{Rco}_{\mathrm{pLS20}}$ and to understand how it may interact with $\mathrm{Rap}_{\mathrm{pLS20}}$ to drive the regulation of the conjugation in the *B. subtilis* pLS20 conjugative plasmid, we embarked on the structure determination of the protein. After several efforts to obtain the structure of the FL protein and the structure of the protein together with its DNA, we were only able to grow crystals that obtained a degradation product, consisting of residues 124-158. $\mathrm{Rco}_{\mathrm{pLS20}}$[124-158] crystals started to appear after 3 months of incubation at conditions described in §C 1.3.1 "Crystallization of $\mathrm{Rco}_{\mathrm{pLS20}}$" and had different shapes, some examples are shown in **FIGURE 12**. All crystals diffracted belonged to $P\ 1\ 2_1\ 1$ space group with a unit cell of a=35.314, b=36.170, c=109.52 and α=90, β=90.179, γ=90 and diffracted to a resolution of up to 1.458Å. The structure contains eight monomers in the asymmetric unit, forming two crystallographically independent tetramers. The X-ray diffraction data and the final structure validation parameters of $\mathrm{Rco}_{\mathrm{pLS20}}$[124-158] are summarized in TABLE **1**.

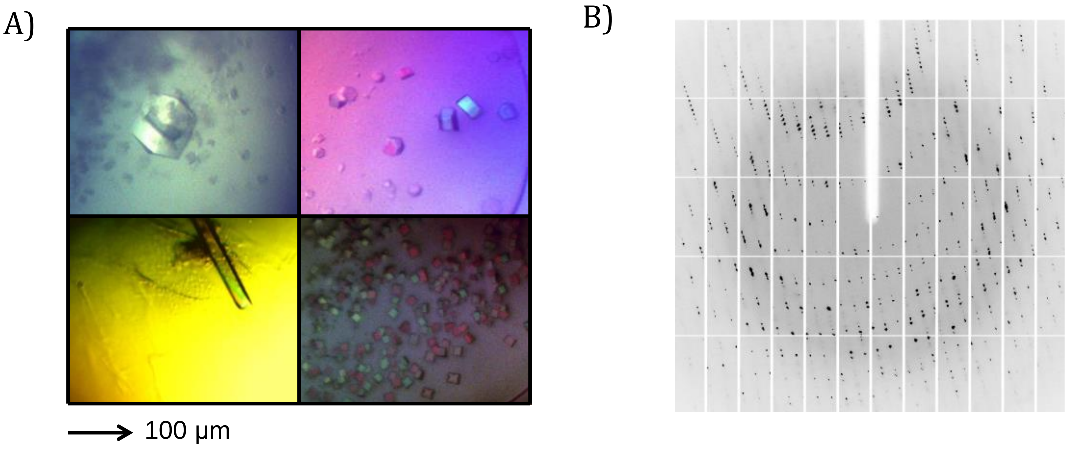


FIGURE 12: **$\mathrm{Rco}_{\mathrm{pLS20}}$[124-158] crystals and its X-ray diffraction pattern. A)** Some examples of $\mathrm{Rco}_{\mathrm{pLS20}}$[124-158] crystals under the microscope. Several different-shaped crystals were obtained after an incubation of approximately 3 months. Pictures were taken with an optical microscope using polarized light. Images are shown in scale. **B)** X-ray diffraction from the crystal of $\mathrm{Rco}_{\mathrm{pLS20}}$[124-158] produced this interference pattern among others. Crystals diffracting up to 1.458Å were obtained.

TABLE 1: X-ray data processing and refinement statistics for Rco$_{pLS20}$[124-158]

Data processing statistics	**Rco$_{pLS20}$[124-158]**
Space group	*P* 1 2_1 1
Unit-cell parameters (Å)	a=36.191 b=35.356 c=109.611
Unit-cell angles (°)	α=90 β=90.179 γ=90
Resolution range (Å)[a]	1.458-36.54
No. of unique reflections	18672
Completeness (spherical) (%)	67.4
Completeness (elliposidal) (%)	90.7
Redundancy	3.5
Mean I/σ(I)	12.8
R_{meas} (%)[b]	0.046
Refinement statistics	
R_{work}[c] (%)	21.03
R_{free}[d] (%)	26.86
Ramachandran	
Favored (%)	98.45
Disallowed (%)	1.55
R.M.S.D.	
Bond lengths (Å)	0.015
Bond angles (°)	1.39
Chirality	
Mean B value (Å^2)	38.0

[a] Numbers in parentheses represent values in the highest resolution shell.

[b] Rmeas=$\sum hkl\ [N/N-1]^{1/2} \sum_i |Ii(hkl) - \langle I(hkl)\rangle| / \sum_{hkl}\sum_i Ii(hkl)$ where N is the multiplicity of a given reflection, Ii(hkl) is the integrated intensity of a given reflection, and <I(hkl)> is the mean intensity of multiple corresponding symmetry-related reflections.

[c] Rwork=$\sum ||Fobs| - |Fcalc|| / \sum |Fobs|$, where |Fobs| and |Fcalc| are the observed and calculated structure factor amplitudes, respectively.

[d] R_{free} is the same as R_{work} but calculated with a 5% subset of all reflections that was never used in refinement.

C 1.4.2. Structure of Rco$_{pLS20}$Tet

As shown in **FIGURE 13**, Rco$_{pLS20}$[124-158] consists on a dimer of primary dimers. It is an independently folding domain whose subunits consist of a short 4AA β-strand, comprising residues R127-D130 and a α-helix, comprising residues E137-K157 linked by a sharp turn, involving residue G133. Two monomers constitute a primary dimer, which is formed by packing of an antiparallel intermolecular β-sheet and antiparallel α-helix. This dimer is stabilized by hydrophobic interactions between F129 from the α-helix of one monomer with the I139 from the β-strand of the other monomer. The second dimerization interface is formed by interaction of the two helices of each dimer with the corresponding helices of the second dimer, forming a four-helix bundle. Residues taking part into the stabilization of the dimer of primary dimers are R141 and E145 from each α-helix, forming four hydrogen bonds. On the extremes of the cluster, the hydrophobic residue L148 is located. Hence, the tetramerization interface is composed of a central cluster of charged residues capped on both sides by L148.

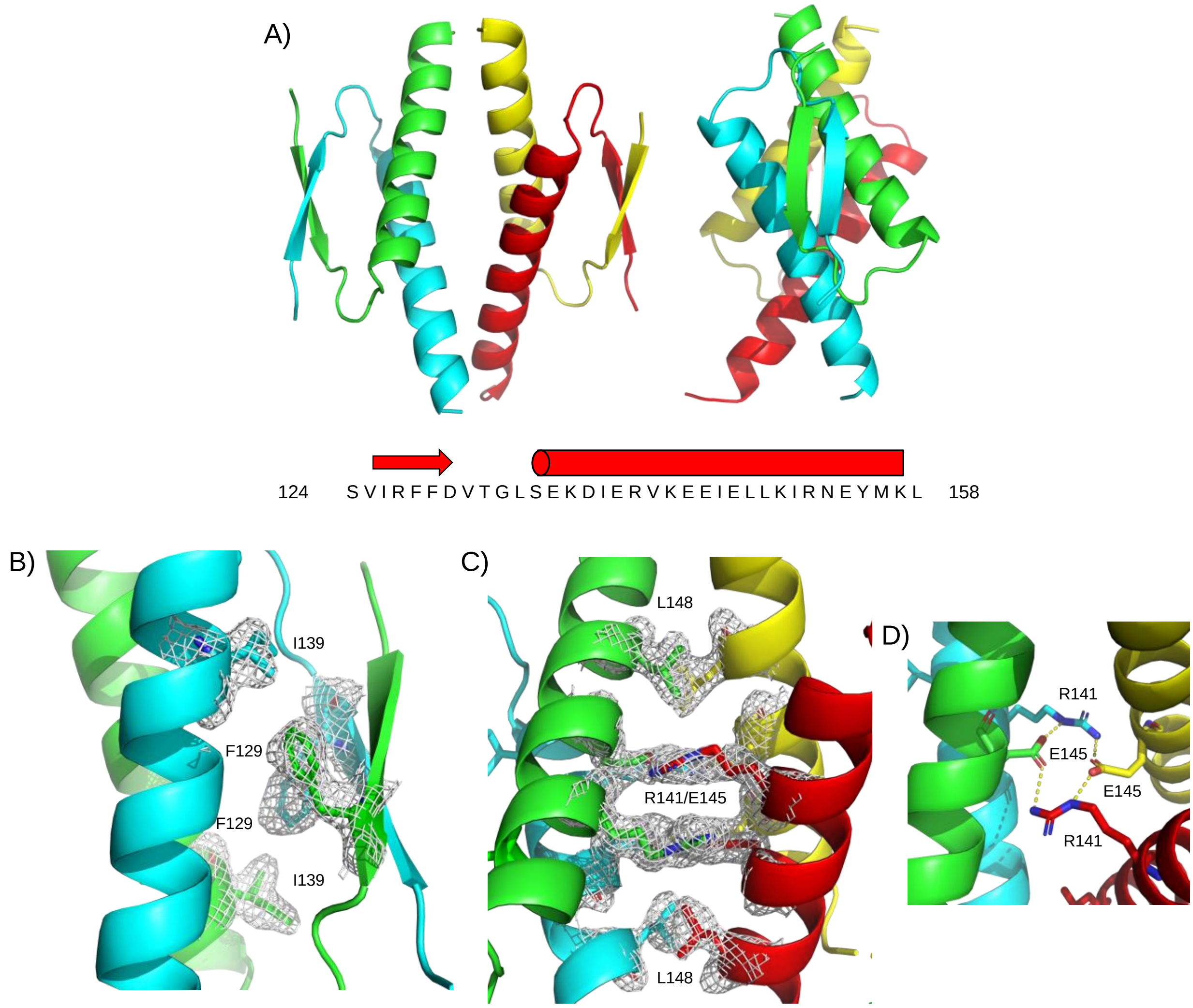

A)
124
S V I R F F D V T G L S E K D I E R V K E E I E L L K I R N E Y M K L
158
B)
I139
F129
F129
I139
C)
L148
R141/E145
L148
D)
R141
E145
E145
R141

FIGURE 13: **Overview of Rco$_{pLS20}$[124-158] structure. A)** Cartoon representation of the dimer of dimers, shown from two different perspectives. The lower panel shows the structurally characterized sequence and the secondary structure. **B)** Close-up view of the hydrophobic stabilization core of the primary dimers, showing the hydrophobic residues F129 and I139 as sticks. **C)** Side view of the polar tetramerization interface. The charged residues R141 and E145 and the L148 cap are represented as sticks. **D)** Representation of the hydrogen bonds formed between the charged residues in the tetramerization interface. All figures are colored by chain and in B) and C) F$_o$-F$_c$ maps are displayed.

Based on the retrieved results, we can conclude that the part of Rco$_{pLS20}$ that we have structurally characterized, encompassing residues 124 to 158, is in charge of the tetramerization and thus we will refer to it as Rco $_{pLS20}$Tet from now on.

C 1.4.3. pH-dependent oligomerization behavior of Rco$_{pLS20}$

Given the charged nature of the dimer-dimer interface of Rco$_{pLS20}$Tet, we decided to probe the oligomerization behavior of Rco$_{pLS20}$ by performing some analytical SEC assays at acidic and basic pHs. In case of Rco$_{pLS20}$Tet, interactions are formed by glutamates and arginines. Under physiological pHs, amino groups are protonated while carboxyl group are not. By changing the pH, the protonation state of the glutamates and arginines that form the dimer-dimer interface will also change. In general terms, when the pH is low, an excess of protons is present and thus, the concentration of protonated molecules is larger than the deprotonated ones. On the contrary, at high pHs, the concentration of deprotonated forms will be larger than the protonated ones.

Therefore, at physiological pHs salt bridges are favoured because of a mixture of protonated and deprotonated species. However, at low and high pHs glutamates and arginines will be protonated and deprotonated, respectively, which may cause a change in the tetramerization interface and thus, in the oligomerization state of Rco$_{pLS0}$. To test this hypothesis, Rco$_{pLS20}$ was injected into a Superdex 200 column at pH 5, pH 8 and pH 10. Chromatograms are displayed at **FIGURE 14**.

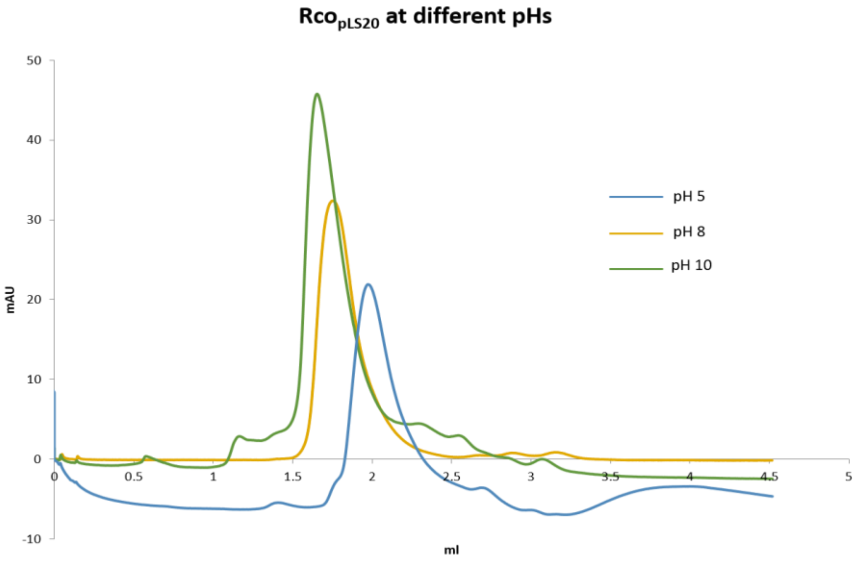


FIGURE 14: **SEC of Rco$_{pLS20}$ by SEC at pH5, pH8 and pH10**. Rco$_{pLS20}$ elution profile at pH5, pH8 and pH10 are represented in blue, orange and green, respectively. The absorbance at 280nm is plotted against V$_{el}$.

TABLE 2: **Summary of the results obtained by analytical SEC for Rco_{pLS20} at different pHs.** The table shows the V_{el} of Rco_{pLS20} at pH5, pH8 and pH10, together with an estimate of M_W and predicted oligomerization state. The M_w are estimated using the calibration of the column. The oligomerization states are calculated taking into acoount that the M_w of Rco_{pLS20} M_W is 20.32 kDa.

Protein/complex	**V_{el} (ml)**	**Estimated M_W based on V_{el} (kDa)**	**Estimated oligomerization state**	**Predicted oligomerization state in solution**
pH 5	1.97	56.99	2.8	Trimer
pH 8	1.80	85.43	4.2	Tetramer
pH 10	1.66	151.57	7.5	Octamer

As we can see in **FIGURE 14** and TABLE **2**, the V_{el} obtained in all cases was remarkably different, confirming the pH has an effect on the oligomerization state of Rco_{pLS20}. We found that Rco_{pLS20} tends to form higher order oligomers at alkaline pHs. At acidic pH, a disruption of the tetramer is observed. This is probably due to the fact that the stabilized hydrogen bonds between R141 and E145 in the tetramerization interface are disrupted by protonation-induce neutralization of the carboxylates of the central E145.

The M_W of the species that form at pH 5, pH 8 and pH 10 were estimated by calibration of the SEC elution column based on the individual elution of a set of proteins of distinct M_W under equivalent conditions. The M_W estimates are 56.99 kDa at pH 5, 85.43 kDa at pH 8 and 151.57 kDa at pH 10, corresponding to 2-3 protein molecules, 4 protein molecules and 8 protein molecules, respectively. Thus, the protein forms tetramers under neutral pH, but has the tendency to form higher order oligomers with increasing pH and at increasing concentrations.[195] We show here that the protein has a preference for formation of octamers at pH 10. No direct evidence for hexamers, heptamers or complexes larger than octamers was observed by SEC in the tested concentrations.

The pH dependence of octamerization shows that this interface is also charged like that of the tetramer. Analysis of the distribution of charges and hydrophobic patches across the surface of the Rco_{pLS20}Tet shows that the face of the hexagonal form of the tetramer is charged, whereas the edges are hydrophobic. Thus, octamerization is likely to be mediated by through face-to-face interactions between Rco_{pLS20}Tet.

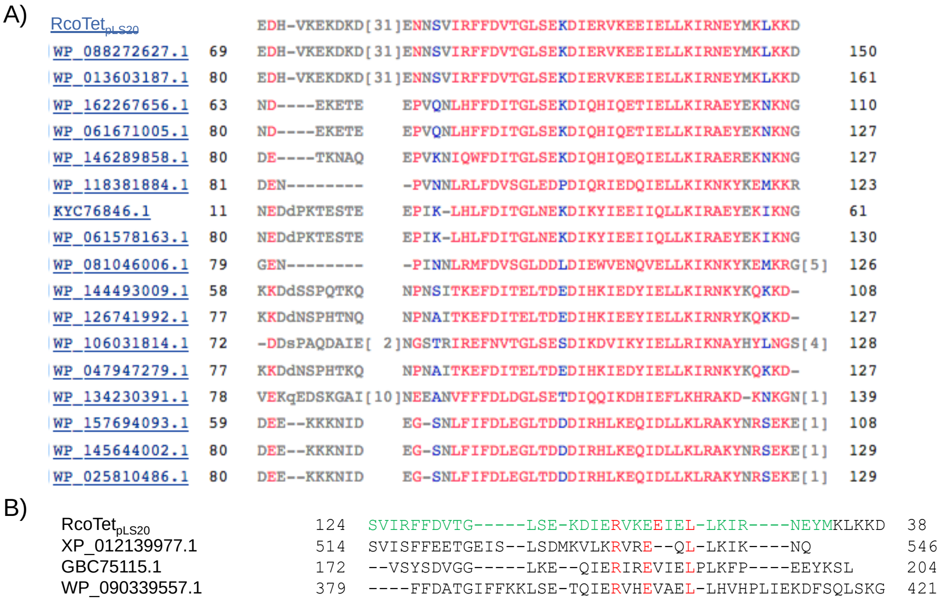
C 1.4.4. Sequence alignments of $\mathrm{Rco_{pLS20}}$Tet

Despite the low sequence homology, we conducted an extensive search for the newly-discovered tetramerization domain of $\mathrm{Rco_{pLS20}}$. As expected, none of the above described structural analogs were identified. Most of the results obtained belonged to the phylum Firmicutes, with 18 sequences with E-values ranging below 1. (**FIGURE 15A**)

Regarding eukaryotes, two uncharacterized proteins with low sequence homology were identified. The first, jerky protein homolog-like isoform X2 from *Megachile rotundata* (accession number: XP_012139977), which contains a C-terminal sequence, comprising residues S514 to Q546, with very low homology with the β-strand and α-helix of $\mathrm{Rco_{pLS20}}$Tet. This protein is annotated to belong to a family of endonuclease-like proteins that have lost their capacity to bind metals. Remarkably, the positions of R141 and E145 are preserved in the alignment.

The second eukaryote protein with sequence homology to $\mathrm{Rco_{pLS20}}$Tet is a protein marked as proteasome-associated ATPase, from archaeon HR06 (accession number: GBC75115.1) encompassing residues V172 to L204. This protein preserves the relative position of the $\mathrm{Rco_{pLS20}}$ R141 and L148 residues, and has a glutamate at position +3 from the arginine, one position off with respect to the +4 position observed for $\mathrm{Rco_{pLS20}}$Tet. Furthermore, the start of the sequence contains a stretch of residues prone to form β-sheets, and the following loop region contains two glycine residues.

A match was also found for one Gram - protein, an ATP-binding cassette domain-containing protein from *Dyadobacter koreensis* (accession number: WP_090339557), residues F379 to G421. Also, the glycine and relative positions of the REL triad are preserved, as are the residues of the β-sheet preceding the loop and helix. (**FIGURE 15B**)



A)

```
RcoTetpLS20             EDH-VKEKDKD[31]ENNSVIRFFDVTGLSEKDIERVKEEIELLKIRNEYMKLKKD
WP_088272627.1   69     EDH-VKEKDKD[31]ENNSVIRFFDVTGLSEKDIERVKEEIELLKIRNEYMKLKKD      150
WP_013603187.1   80     EDH-VKEKDKD[31]ENNSVIRFFDVTGLSEKDIERVKEEIELLKIRNEYMKLKKD      161
WP_162267656.1   63     ND----EKETE    EPVQNLHFFDITGLSEKDIQHIQETIELLKIRAEYEKNKNG      110
WP_061671005.1   80     ND----EKETE    EPVQNLHFFDITGLSEKDIQHIQETIELLKIRAEYEKNKNG      127
WP_146289858.1   80     DE----TKNAQ    EPVKNIQWFDITGLSEKDIQHIQEQIELLKIRAEREKNKNG      127
WP_118381884.1   81     DEN--------    -PVNNLRLFDVSGLEDPDIQRIEDQIELLKIKNKYKEMKKR      123
KYC76846.1       11     NEDdPKTESTE    EPIK-LHLFDITGLNEKDIKYIEEIIQLLKIRAEYEKIKNG      61
WP_061578163.1   80     NEDdPKTESTE    EPIK-LHLFDITGLNEKDIKYIEEIIQLLKIRAEYEKIKNG      130
WP_081046006.1   79     GEN--------    -PINNLRMFDVSGLDDLDIEWVENQVELLKIKNKYKEMKRG[5]   126
WP_144493009.1   58     KKDdSSPQTKQ    NPNSITKEFDITELTDEDIHKIEDYIELLKIRNKYKQKKD-      108
WP_126741992.1   77     KKDdNSPHTNQ    NPNAITKEFDITELTDEDIHKIEEYIELLKIRNRYKQKKD-      127
WP_106031814.1   72     -DDsPAQDAIE[ 2]NGSTRIREFNVTGLSESDIKDVIKYIELLRIKNAYHYLNGS[4]   128
WP_047947279.1   77     KKDdNSPHTKQ    NPNAITKEFDITELTDEDIHKIEDYIELLKIRNKYKQKKD-      127
WP_134230391.1   78     VEKqEDSKGAI[10]NEEANVFFFDLDGLSETDIQQIKDHIEFLKHRAKD-KNKGN[1]   139
WP_157694093.1   59     DEE--KKKNID    EG-SNLFIFDLEGLTDDDIRHLKEQIDLLKLRAKYNRSEKE[1]   108
WP_145644002.1   80     DEE--KKKNID    EG-SNLFIFDLEGLTDDDIRHLKEQIDLLKLRAKYNRSEKE[1]   129
WP_025810486.1   80     DEE--KKKNID    EG-SNLFIFDLEGLTDDDIRHLKEQIDLLKLRAKYNRSEKE[1]   129
```

B)

```
RcoTetpLS20         124   SVIRFFDVTG-----LSE-KDIERVKEEIEL-LKIR----NEYMKLKKD   38
XP_012139977.1      514   SVISFFEETGEIS--LSDMKVLKRVRE--QL-LKIK----NQ          546
GBC75115.1          172   --VSYSDVGG-----LKE--QIERIREVIELPLKFP----EEYKSL      204
WP_090339557.1      379   ----FFDATGIFFKKLSE-TQIERVHEVAEL-LHVHPLIEKDFSQLSKG   421
```

FIGURE 15: **Multiple sequence alignments (MSA) obtained by BLAST search of $\mathrm{Rco_{pLS20}}$Tet.** **A)** Sequence alignment of results obtained in the Firmicutes phylum with an E<1. **B)** Combined sequence alignment of results obtained in eukaryotes and Gram – bacteria. Green-colored part corresponds to structurally characterized part of $\mathrm{Rco_{pLS20}}$. REL triad is represented in red.

C 1.4.5. Folding comparison of Rco$_{pLS20}$Tet

After having elucidated the structure of the domain that corresponds to residues 124-158 of Rco$_{pLS20}$, we looked for structural homologs in the PDB database[202], and so we launched PDBeFOLD (EMBL-EBI). The 20 most similar hits are shown in **TABLE 3**.

TABLE 3: **Structure alignment results for Rco$_{pLS20}$Tet.** Different scores used for structure recognition are shown together with the protein accession number. They all concern p53 protein family members belonging to *Homo sapiens*. Results are sorted by Q-score and only the twenty most relevant ones are presented.

##	Scoring			RMSD	Nalign	N	%seq	Query			Protein accession number
	Q	P	Z					%sse	Match	Nnes	
1	0.73	2.7	4.7	1.46	29	1	14	100	2wqj:I	29	O15350
2	0.72	2.4	4.4	1.09	26	1	15	100	2wqj:B	26	O15350
3	0.72	2.3	4.3	1.41	28	1	14	100	3zy0:C	28	Q9H3D4
4	0.71	2.4	4.4	1.33	28	1	14	100	3zy0:A	29	Q9H3D4
5	0.71	2.1	4.0	1.47	28	1	14	100	2qwj:U	28	O15350
6	0.68	2.3	4.3	1.32	26	1	15	100	3zy0:D	26	Q9H3D4
7	0.67	2.1	4.1	1.78	30	1	17	100	2j10:A	31	P04637
8	0.65	1.9	3.8	1.63	29	1	17	100	2j10:C	31	P04637
9	0.65	2.2	4.2	1.86	30	1	13	100	2j0z:A	31	P04637
10	0.65	2.1	4.1	1.63	29	1	17	100	2j10:B	31	P04637
11	0.65	2.3	4.3	1.65	29	1	14	100	2j0z:D	31	P04637
12	0.65	2.1	4.1	1.89	30	1	13	100	2j0z:B	31	P04637
13	0.65	2.4	4.3	1.90	30	1	17	100	2j10:D	31	P04637
14	0.64	2.1	4.1	1.69	29	1	14	100	2j11:D	31	P04637
15	0.63	2.3	4.3	1.18	26	1	15	100	2wqj:A	29	O15350
16	0.63	2.1	4.0	1.76	29	1	14	100	2j0z:C	31	P04637
17	0.61	2.1	4.1	1.85	29	1	14	100	2j11:C	31	P04637
18	0.52	3.3	5.2	1.60	30	1	13	100	1saf:A	42	P04637
19	0.50	2.7	4.7	1.73	30	1	13	100	1olh:C	42	P04637
20	0.50	2.2	4.2	1.54	29	2	14	100	1saf:B	42	P04637

By running PDBeFOLD, 113 hits were obtained (in **TABLE 3** only the 20 most relevant ones are shown). From these hits, 77 belonged to *Homo sapiens*. More interestingly, 74 out of 77 *H. sapiens* hits were components of the p53 family (either p53, p63 or p73).

We also examined PDBeFOLD results based on the function of the hits obtained, rather than the organism they belonged to. Among the 113 different PDB entries, 88 concerned ODs. Within the ODs, there were both tetramerization and dimerization domains. Mainly, the dimerization domains were obtained by introducing double mutations in originally tetramerization domains (for instance, PDB code: 1HS5), by changing amino acids with hydrophobic side chains by electrically charged ones.[203]

Thus, several significant hits were found, yet most of them corresponded to the tetramerization domain of p53 family proteins (**TABLE 3**). The basic fold of all of these structures consists of the short β-sheet followed by an α-helix, connected by a loop. Both β-sheet and α-helix have very similar lengths and structural differences are mainly found in the angle between the β-sheet and the α-helix. Even if the RMSD are of about 1.2-1.8Å, low sequence similarity is observed with the hits obtained with a maximum of a 17% of identity. Apparently, structural and sequence alignments do not coincide. (**FIGURE 16**)

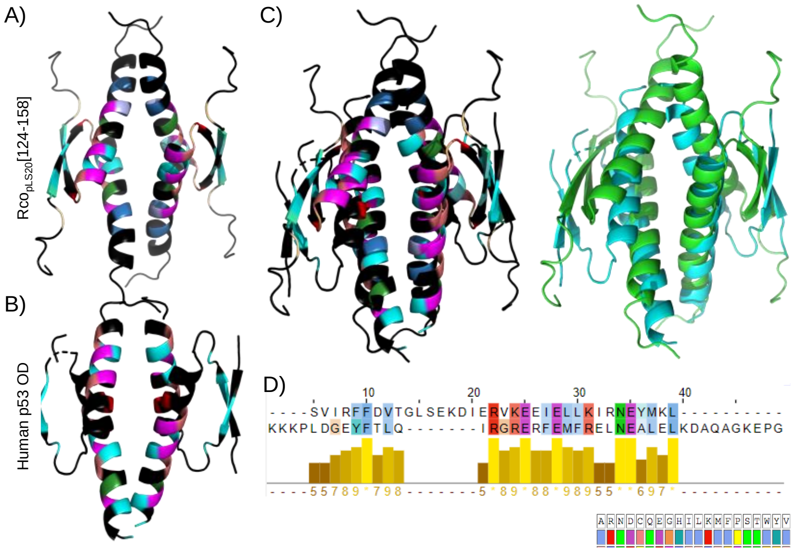

A)
C)
Rco$_{pLS20}$[124-158]
B)
Human p53 OD
D)
10
20
30
40
- - - - S V I R F F D V T G L S E K D I E R V K E E I E L L K I R N E Y M K L - - - - - - - - - - -
K K K P L D G E Y F T L Q - - - - - - - - I R G R E R F E M F R E L N E A L E L K D A Q A G K E P G
- - - - 5 5 7 8 9 * 7 9 8 - - - - - - - - 5 * 8 9 * 8 8 * 9 8 9 5 5 * * 6 9 7 * - - - - - - - - - - -
A R N D C Q E G H I L K M F P S T W Y V

FIGURE 16: **Comparison of Rco$_{pLS20}$Tet and human p53 OD (PDB code: 1SAE). A,B)** Rco$_{pLS20}$Tet structure (A) and its homolog human p53 OD structure (B). Residues that are conserved in the sequence alignment are represented in different colors. **C)** Structural superposition of both domains. Superposition with conserved residues from sequence alignment and a plain superposition are shown, in which Rco$_{pLS20}$Tet is represented in green and human p53 OD in blue. **D)** Sequence alignment of both domains together with conservation scores. Color legend for each residue is displayed at the bottom right corner.

It is worth mentioning that due to the huge impact cancer has in our society and given the importance of p53, p63, p73 proteins in this disease, there is a large number of research groups working on this matter. Consequently, PDB entries and results in this field are also abundant. There is a possibility that Rco$_{pLS20}$Tet has structural similarities with some evolutionary closer relatives but we were unable to identify them by the 3D alignment of protein structures probably because the database is incomplete and results are slanted. In any case, the RMSD scores are importantly low, virtually reaching 1.2Å at some cases, which leads us to believe that the resemblance between Rco$_{pLS20}$Tet and the p53 family members is not arbitrary.

C 1.4.6. Comparison between Rco$_{pLS20}$Tet and P53 family proteins

The most relevant and surprising finding after the PDBeFold analysis was the fact that Rco$_{pLS20}$ has an OD that shares a high degree of similarity with that of vertebrate p53 family members. P53 is a member of a family that includes the homologs p63 and p73. They are the core of numerous signaling pathways that determine the fate of the cell.[204,205] Each member of the family has specific functions during embryonic development. p63 and p73 have several isoforms that have divergent and even antagonist functions.[206] DNA binding domains of p63 and p73 share remarkable homology to those of p53, therefore, they upregulate a number of the same of downstream target genes.[207]

Stress stimuli trigger all three members of p53 family to act as transcription factors. They are translocated from the cytoplasm to the nucleus and drive the expression of transcripts involved in regulating cell cycle arrest and apoptosis.[208,209] Moreover, the capability of p53 to stimulate apoptosis seems to be dependent on the activation of p63 and p73.[210,211] In case of Rco$_{pLS20}$, as explained in the introductory part, it is expressed in low amounts as conjugation it is a highly time-consuming process and not always is needed. On the contrary, p53 family members may be expressed by default as the genes they regulate are involved in processes that are crucial for cells' survival. However, what these two mechanisms have in common is the need for strict regulation. One of the mechanisms to achieve a proper transcription regulation is via the formation of a DNA loop, for which high-order oligomers like tetramers are necessary.

Since its discovery in 1979, p53 has been on the cutting edge of cancer research. Mutations in p53 are the most prevalent genetic alteration in human cancer.[212,213] Also, studies have demonstrated that p53 expression correlates with tumor grade and recurrence.[214] The identification and characterization of p53 mutations led to the description of seven mutational hotspots, which are often found in tumors.[215] The phylogenetically more ancient family members p63 and p73 were initially thought to have a developmental role.[216,217] However, we now have evidence that p63 and p73 can also act as tumor suppressors. Depending on the cellular context, they can exert their function as tumor suppressors together with p53 or in and independent manner.[204,218]

Oligomerization is necessary for p53 to function[219], which is a logical consequence of the observation that p53 binds to DNA as a tetramer.[220,221] Besides the OD in p53, it also contains a multifunctional C-terminal domain (CTD), which is highly basic and intrinsically disordered and participates in many aspects of p53 functions, such as recruitment of co-factors, regulation of protein stability or its binding behavior to DNA.[222] It also contains a N-terminal transactivation domain (TA) for recruiting transcriptional factors and a DNA binding domain (DBD).

The most significant difference of p63 and p73 compared to p53 is the presence of additional structural elements in the C-terminal region of the protein. They both contain an extended C-terminal region, which also includes a sterile alpha motif (SAM), which is a small protein–protein interaction module that is found in a wide variety of different proteins, ranging from kinases and transcriptional regulators to cell surface receptors.[223] Furthermore, p63 also contains a transaction inhibitory domain (TID) that plays a role in stabilizing a latent dimeric form.[224] The fact that p53 lacks additional stabilizing structural elements suggests that it is the most recent evolutionary member.

The experiments described in this chapter show that Rco$_{\mathrm{pLS20}}$Tet structure folding is highly similar to that of p53 family members, also resulting in dimer of dimers that generate a four-helix bundle. Rco$_{\mathrm{pLS20}}$Tet comprises residues 124-158, which excludes an uncharacterized C-terminal part that could be a α-helix as is the case in p63 and p73. Based on secondary structure predictions by PSIPRED 4.0 (UCL-CS Bioinformatics),[225,226] we predict that a 10 residue long sequence comprising residues 161 to 172 in the C-terminus is likely to form an α-helix, highlighted in **FIGURE 17**. Remarkably, the α-helix structurally characterized is predicted but not the β-sheet.

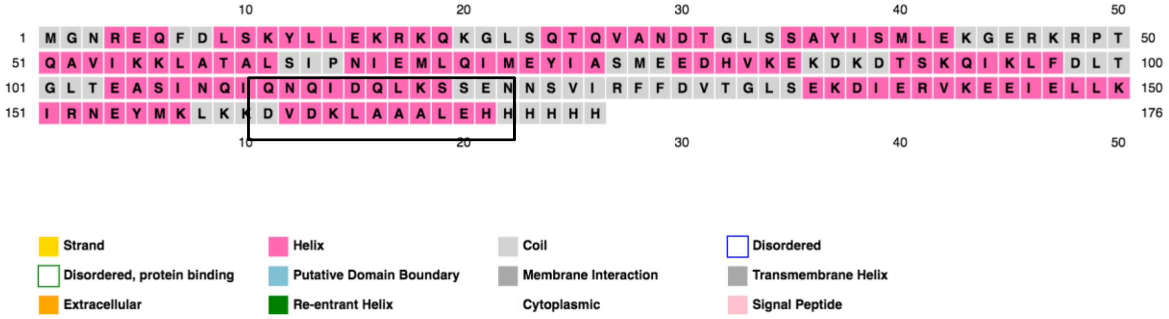

10 20 30 40 50
1 M G N R E Q F D L S K Y L L E K R K Q K G L S Q T Q V A N D T G L S S A Y I S M L E K G E R K R P T 50
51 Q A V I K K L A T A L S I P N I E M L Q I M E Y I A S M E E D H V K E K D K D T S K Q I K L F D L T 100
101 G L T E A S I N Q I Q N Q I D Q L K S S E N N S V I R F F D V T G L S E K D I E R V K E E I E L L K 150
151 I R N E Y M K L K K D V D K L A A A L E H H H H H H 176
10 20 30 40 50
Strand
Helix
Coil
Disordered
Disordered, protein binding
Putative Domain Boundary
Membrane Interaction
Transmembrane Helix
Extracellular
Re-entrant Helix
Cytoplasmic
Signal Peptide

FIGURE 17: **Secondary structure prediction of Rco$_{\mathrm{pLS20}}$.** Rco$_{\mathrm{pLS20}}$ FL sequence was entered and the result is displayed by colors. The potential extra α-helix is highlighted by a black square. Color legend is shown below.

There are structures available for eukaryote, non-human p53 homologs, which include *Drosophila melanogaster* (Dmp53) and *Caenorhabditis elegans* (CEP-1). These diverged long ago from the vertebrates and do not possess exactly the same quaternary structure in their ODs (**FIGURE 20**) Consistently, they cannot be identified as homologs by sequence analysis. Interestingly, they show a similar DNA binding specificity. In fact, the only sequence homology between Dmp53 and CEP-1 with p53 family members is confined to the DBD.[227,228]

Even if key residues in all vertebrate p53 ODs are strictly conserved, this is not the case in the OD of CEP-1 and Dmp53. P53 OD contains a turn between the β-strand and α-helix enabled by a glycine residue that is highly conserved across many different species. However, CEP-1 and Dmp53 contain a serine, threonine and a proline, asparagine dipeptide, respectively, in this position. In case of Rco$_{\mathrm{pLS20}}$Tet, we have realized that a glycine residue is also present in the sharp turn (G133), which reveals another key element that is maintained in regards to p53 family proteins OD besides the folding. (**FIGURE 18**)

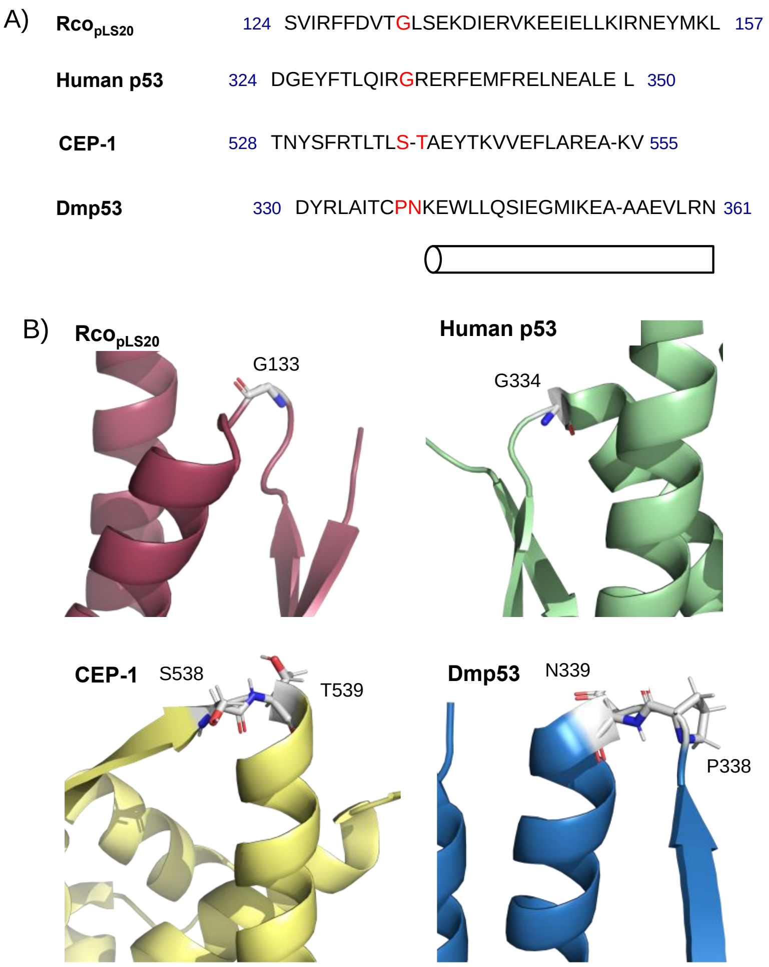

A)
Rco$_{pLS20}$ 124 SVIRFFDVTGLSEKDIERVKEEIELLKIRNEYMKL 157
Human p53 324 DGEYFTLQIRGRERFEMFRELNEALE L 350
CEP-1 528 TNYSFRTLTLS-TAEYTKVVEFLAREA-KV 555
Dmp53 330 DYRLAITCPNKEWLLQSIEGMIKEA-AAEVLRN 361
B)
Rco$_{pLS20}$
G133
Human p53
G334
CEP-1
S538
T539
Dmp53
N339
P338

FIGURE 18: **Sequence divergence in the OD of Rco$_{pLS20}$, human p53, CEP-1 and Dmp53. A)** Structure alignment of the glycine in the sharp turn between the β-strand and the α-helix is conserved in Rco$_{pLS20}$ while it is not in CEP-1 and Dmp53. Glycine and the residues replacing it are represented in red. The glycine residue is also present in all p53, p63 and p73 proteins of other vertebrates. **B)** Cartoon representation showing the location of the glycine in Rco$_{pLS20}$ and human p53 or dipeptide residues in CEP-1 and Dmp53. Glycine and dipeptides are displayed as sticks.

Importantly, we have realized there is a correlation between the presence of the glycine in the loop with the widening of the angle between the α-helix and the β-sheet. For the glycine containing structures, the largest angle is 27°, whereas for Dmp53 and CEP-1, which contain dipeptides instead of the glycine, the angles are 34° and 62°, respectively. (**FIGURE 19**)

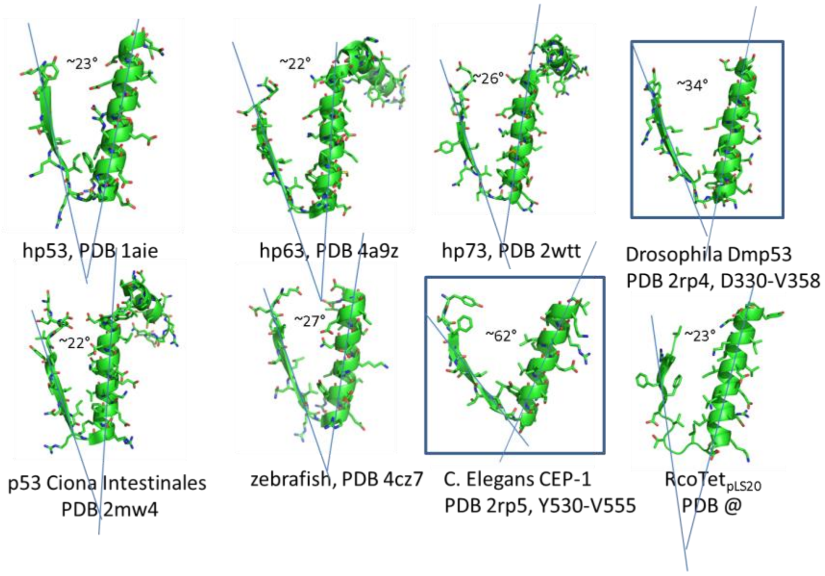


FIGURE 19: **Cartoon and stick representation of the core sheet-loop-helix motif of the $\mathrm{Rco}_{\mathrm{pLS20}}$Tet-like structures.** The angles between the α-helix and the β-sheet are shown for different p53 homologs. CEP-1 and Dmp53 are highlighted by a square box as they do not contain the conserved glycine residue.

Another major difference found in CEP-1 compared to p53 family members is related to the oligomerization state and the tetramerization interface. The classic p53 protein family OD is a dimer of dimers, involving two separate and distinct interfaces.[229] This quaternary structure is maintained in $\mathrm{RcoTet}_{\mathrm{pLS20}}$. Nevertheless, CEP-1 forms dimers.[230]

The first dimer interface, which involves contacts between the antiparallel β-sheet and α-helix, is stabilized by hydrophobic interactions in all p53 protein family members. In case of p53, the second dimerization interface is also composed of hydrophobic interactions between the four α-helices of both dimers. This is achieved by hydrophobic interactions between M340, L344, A347, L348 and L350 from each α-helix. In contrast, the equivalent residues of CEP-1 are the charged residues K544, R551 and E552. The lack of charge complementarity between residues reveals the reason why CEP-1 forms dimers instead of tetramers.[230]

In case of Dmp53, it has extra residues which lead to a different fold that cannot be found in any p53 proteins, in which an extra β-strand and an extra α-helix are present compared to the canonical p53 OD fold. In this case, the tetramerization interface consists of the inner four helices from the four monomers and it is stabilized by a central cluster of charged residues that form salt bridges between K352 and E353.[230]

By studying $\mathrm{Rco}_{\mathrm{pLS20}}$Tet structure we see how its arrangement is similar to that of p53 protein family members or CEP-1, composed of a dimer of dimers. However, the central core is mainly composed of charged residues (R141, E145) instead of hydrophobic ones, as it is in Dmp53. How hydrogen bonds are formed between R141 and E145 residues is represented in (**FIGURE 13D**). The overall structure of $\mathrm{Rco}_{\mathrm{pLS20}}$Tet, human p53, CEP-1 and Dmp53 and the residues involved in their tetramerization interfaces are represented in **FIGURE 20**.

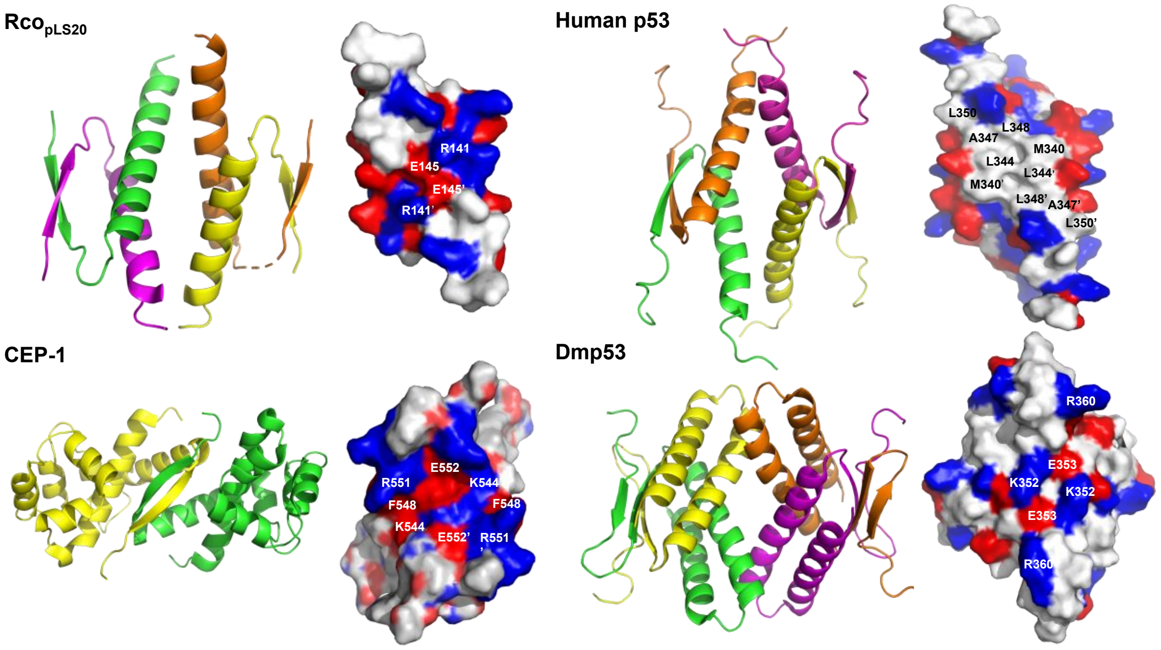


FIGURE 20: **Cartoon and surface representations of the tetrameric interface of the OD of Rco$_{pLS20}$, p53, CEP-1 and Dmp53.** In Rco$_{pLS20}$Tet, R141 and E145 from one dimer form salt bridges with R141 and E145 from the other dimer. In case of p53, the central core is composed of M340, L344, A347, and L348, making the tetrameric core mainly hydrophobic. CEP-1 is composed of charged residues K544, R551 and E552 that repel themselves. The tetrameric interface of Dmp53 is composed of charged residues K352 and E353 forming salt bridges as in Rco$_{pLS20}$Tet with equivalent residues being R141 and E145. In all surface representations the other dimer is not shown. Positively charged residues are shown in blue, negatively charged in red and non-polar ones in white.

As far as functionality is concerned, both Dmp53 and CEP-1 induce apoptosis, which suggests that cell cycle arrest prompted by p53 family members was a later evolutionary development. Of course Rco$_{pLS20}$Tet does not induce apoptosis, but it shares the tetrameric configuration with p53 homologs, probably related to the transcriptional regulation via DNA looping.

Thus, compared to p53 homologs, Rco$_{pLS20}$Tet is structurally very similar to human p53. However, the dimer interface is stabilized by hydrophobic interactions in case of p53, which does not occur in Rco$_{pLS20}$Tet, as the stabilization of the interface involves charged residues. This is a unique feature that has only been described in Dmp53 from all p53 homologs up to date. Thus, Rco$_{pLS20}$Tet seems to have an identical overall structure of p53 but contains a charged core forming salt bridges to stabilize the dimer of dimers.

Results obtained from structural alignment indicate that the Rco$_{pLS20}$Tet region we structurally characterized is highly similar to p53 protein family tetramerization domain despite low sequence similarity.

Tetramerization and DNA looping are intimately linked, both in prokaryotic and eukaryotic systems, despite the lack of structural homology between the proteins forming the links. By these results we confirm the relevance of tetramerization of DNA looping for the *B. subtilis* pLS20 conjugation operon repressor, $\text{Rco}_{\text{pLS20}}$, confirmed by gel electrophoresis analysis by Ramachandran *et al.*[180] However, the structural basis of this mechanism was not understood at that time. The DNA binding region was readily identified as a HTH motif at the N-terminus. However, the low sequence homology with other proteins of known function hampered the identification of the domain responsible for the looping activity. By unraveling the structure of $\text{Rco}_{\text{pLS20}}$Tet we found out that it has structural homology with p53/p63/p73 family of tumor repressor that previously had been observed only in eukaryotes.

Despite the low sequence homology we have been able to propose a motif for the tetramerization domain. The glycine in the sharp turn between the α-helix and β-sheet may be important as it is conserved in metazoan homologs of human p53. Thus, we have found the sheet-kink-helix motif to be present in four proteins. However, the short sequence of the tetramerization domain hampers identification of additional members, thus, we infer from the occurrence of this domain across many domains of life that many others must exist. Nevertheless, the amino acids responsible for the kink and for tetramerization have changed significantly, while preserving the fold of the monomers and the oligomerization behavior.

One of the questions that remains unanswered is whether the four DBDs of $\text{Rco}_{\text{pLS20}}$ bind DNA strand on one end of the loop, closing the loop by forming an octamer stabilized by interactions between the tetramers, or whether the DBDs of a single tetramer bind REs across the DNA loop, where the whole structure is stabilized by cooperativity between additional tetramers binding across the loop. Secondary structure prediction of the residues preceding the OD show that there are no long flexible loops that allow the DBDs to reach distant binding sites. Thus, the distribution of the DBDs around the $\text{Rco}_{\text{pLS20}}$Tet structure matches the across-loop model better.

The model that emerges from the across-loop binding of the tetramer in conjunction with the higher order complexes under proton depletion, is that of cooperative binding of recognition sites across the DNA loop. The initial $\text{Rco}_{\text{pLS20}}$Tet-DNA complex would then be formed stochastically, perhaps with the aid of helper proteins that shape the DNA at the turn, and through the intrinsic propensity of the DNA to bend at the loop region. Binding of several tetramers across the loop stabilizes the loop structure.

Another question that arises is whether the structural similarity between $\text{Rco}_{\text{pLS20}}$Tet and p53 family proteins has evolved independently or if these proteins share a common ancestor. We do not possess sufficient information to resolve this for now as we have encountered some difficulties trying to find an answer to it for two reasons. On the one hand, tetramerization interfaces have been proven to be stabilized by both charged and hydrophobic residues in different species. Thus, a similar fold can occur for highly divergent sequences. On the other hand, the shortness of the sequences complicates the search for homologs because homology is more difficult to detect for shorter sequences. Our search of homologs across the sequence databases suggests that sequential homologs can be detected in varying species. This supports the divergent model, suggesting that all members of this family stem from a common ancestor, although convergent evolution may have occurred for some of these proteins.

The similarity between structures of Rco_{pLS20}Tet and P53 family proteins OD suggest the fold is older than the occurrence of multicellular life.[231] Therefore, this fold could be ubiquitous assuming that they share a common ancestor. Also, interference of bacterial p53-like repressors with the oncogene p53 should be interrogated. Interference with p53 function is known for viral proteins.[232] Moreover, it is known that prolonged bacterial infections can lead to carcinogenesis[233] and that p53 can play in a role in this process.[232] We believe that this possible interaction is an interesting matter of study for the future.

C 1.5. CONCLUSIONS

i. We have been able to obtain a partial Rco_{pLS20} structure, comprising 30 residues of the C-terminal, which we have named Rco_{pLS20}Tet.

ii. The Rco_{pLS20}Tet structure is composed of a dimer of dimers. Each monomer comprises a short β-strand and an α-helix, linked by a sharp turn. The structure was obtained at resolution of 1.458Å and belongs to $P\ 2_1$ space group.

iii. Rco_{pLS20} is a tetramer in solution but its oligomerization state is altered by the pH. The formation of the tetramer is disrupted at low pHs, while at high pHs octamers are formed. The pH dependence of octamerization shows that this interface is also charged like that of the tetramer.

iv. Surprisingly, structure alignments related the structure obtained with those of p53 family proteins. Alignments between Rco_{pLS20} and human p53 showed a RMSD as low as 1.20Å. A glycine residue that is highly conserved in the sharp turn between α-helix and β-strand in all vertebrate p53 family proteins is also present in Rco_{pLS20}Tet.

v. However, by BLAST searches of Rco_{pLS20}Tet sequence we have not find any hits in *H. sapiens*, indicating low sequence homology between structurally very similar proteins. The hits that have been found concerned bacterial species, especially bacteria belonging to the Firmicutes phylum containing a HTH domain, which is a common feature in regulatory proteins.

vi. We have been able to find four sequence homologs in eukaryotes. All of them posses the REL triad. The shortness of Rco_{pLS20}Tet sequence hampers the search of homologs, meaning there could be more homologs we have not been able to find.

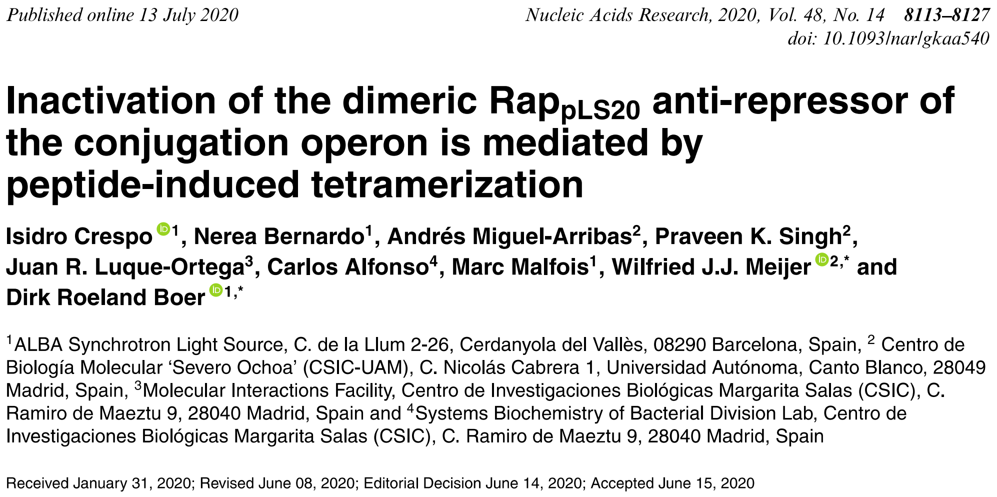
vii. The tetramerization interface of Rco_{pLS20}Tet is stabilized by salt bridges between the amino group of R141 and the carboxyl group of E145. In contrast, tetramerization interface of human p53 is stabilized by hydrophobic interactions. While p53 family proteins and CEP-1 from *C. elegans* are stabilized by hydrophobic interactions, Dmp53 from *D. melanogaster* is also stabilized by salt bridges. Nevertheless, the overall structure of Dmp53 differs from that of Rco_{pLS20}Tet.

viii. Our results suggest the possibility that this fold may be ubiquitous assuming that they share a common ancestor.

CHAPTER 2: Rap$_{pLS20}$, Phr*$_{pLS20}$ and Rco$_{pLS20}$: Regulation of conjugation of pLS20

Part of the results presented in this chapter have been published:

doi: 10.1093/nar/gkaa540



Published online 13 July 2020 *Nucleic Acids Research, 2020, Vol. 48, No. 14* ***8113–8127***
doi: 10.1093/nar/gkaa540

Inactivation of the dimeric Rap$_{\text{pLS20}}$ anti-repressor of the conjugation operon is mediated by peptide-induced tetramerization

Isidro Crespo [1], Nerea Bernardo [1], Andrés Miguel-Arribas [2], Praveen K. Singh [2], Juan R. Luque-Ortega [3], Carlos Alfonso [4], Marc Malfois [1], Wilfried J.J. Meijer [2,*] and Dirk Roeland Boer [1,*]

[1] ALBA Synchrotron Light Source, C. de la Llum 2-26, Cerdanyola del Vallès, 08290 Barcelona, Spain, [2] Centro de Biología Molecular 'Severo Ochoa' (CSIC-UAM), C. Nicolás Cabrera 1, Universidad Autónoma, Canto Blanco, 28049 Madrid, Spain, [3] Molecular Interactions Facility, Centro de Investigaciones Biológicas Margarita Salas (CSIC), C. Ramiro de Maeztu 9, 28040 Madrid, Spain and [4] Systems Biochemistry of Bacterial Division Lab, Centro de Investigaciones Biológicas Margarita Salas (CSIC), C. Ramiro de Maeztu 9, 28040 Madrid, Spain

Received January 31, 2020; Revised June 08, 2020; Editorial Decision June 14, 2020; Accepted June 15, 2020

C 2.1. OBJECTIVES

Rco$_{pLS20}$ and Rap$_{pLS20}$ are key players in the regulation of the genetic switch for activation/deactivation of expression of the conjugative machinery of plasmid pLS20. Rap$_{pLS20}$ apo and Phr*$_{pLS20}$-bound structure have recently been obtained and we now have some structural information about Rco$_{pLS20}$ structure. Still, knowledge on the structural mechanism of regulation of these proteins remains limited. Information about this would be of extreme importance because conjugation would result impaired by inhibiting the binding between Rap$_{pLS20}$ and Rco$_{pLS20}$ and thus, transfer of ARGs would not be possible. Also, Phr*$_{pLS20}$ has been demonstrated to induce Rap$_{pLS20}$ tetramerization. An interesting hypothesis that we have studied in this chapter is the possibility that cognate peptides of other Rap proteins could induce the same effect in Rap$_{pLS20}$, which would mean there could be some cross-regulation between the different systems.

To gain some information on these topics, we established some specific objectives for this chapter:

- Biophysically characterizing the Rco$_{pLS20}$-Rap$_{pLS20}$ complex at different stoichiometries and checking the effect of Phr*$_{pLS20}$ on the complex fomation.
- Analyzing binding between Rap$_{pLS20}$ and cognate peptides from other Rap system to investigate the possibility of cross-regulation between different Rap systems.
- In case we find that any peptide other than Phr*$_{pLS20}$ is able to bind Rap$_{pLS20}$ we plan to attempt to solve the atomic structure of the complex with Rap$_{pLS20}$ to elucidate the contacts formed and compare them to that of Phr*$_{pLS20}$.

C 2.2. INTRODUCTION

In the previous chapter, we outlined how important it is to regulate the conjugation genes and we introduced the main players that take part in this process: Rco$_{pLS20}$, Rap$_{pLS20}$ and Phr*$_{pLS20}$. Generally, genes responsible for conjugation are maintained in a default “OFF” state and are induced by signaling molecules to be activated to an “ON” state. In some systems, conjugation genes are constitutively expressed but limited in their expression. On the one hand, constitutive systems, for instance those of the IncF plasmid family. These plasmids integrate transcriptional cues of plasmid and host factors, together with environmental stimuli that control the expression of their transfer region. However, inducible conjugative systems are based on sensing phenolic compounds, also known as quorum-sensing (QS) systems.

QS is a common way by which bacteria communicate with one another using small and diffusible chemical signaling molecules. When the concentration of a signaling molecule reaches a certain level, bacteria respond by changing their gene expression.[234,235] Several cellular processes in both Gram + and Gram - bacteria are regulated by QS, for instance the development of natural competence in *B. subtilis* and *S. pneumonia*.[192] QS has also been reported to regulate other conjugation genes, like those involved in the transfer of the tumor-inducing pTI plasmid of Gram - bacteria *A. tumefaciens* into plant cells. Among the QS systems occurring in Gram + bacteria, regulation of conjugation of *Enterococci* plasmids pCF10 and pAD1 has been notably studied. The transfer of these plasmids is controlled by the ratio of two peptides, a plasmid-encoded peptide (iCF10/iAD1) inhibits the regulator (TraA/PrgX) by keeping the operon in an "OFF" state, and a chromosomally encoded peptide (cCF10/iAD1) relieves transcriptional repression of the conjugation operon through a competitive binding to the master regulator.

The RRNPP family of Gram + tetratricopeptide (TPR) proteins was named after the discovery of the main representatives of the family (Rap/NprR/PlcR/PrgX).[236,237] In Gram + bacteria they play a vital role in QS as they function as targets for their cognate signaling peptide.[238,239] The peptide is produced, secreted, processed and then reimported into the cell. Generally, this peptide is about 5-10 amino acids long, consisting of a small fragment of the C-terminus of the full-length pre-protein.[192] The processed peptide binds to the C-terminal TPR domain of the RRNPP proteins. This modulates the interaction between the N-terminal part of RRNPP and the effector molecule, conducting downstream effects.[237] The downstream effector molecule can be a protein or a DNA molecule. Therefore, the function of the RRNPP proteins depends on the nature of the effector molecule and on the type of effector domain of the protein. The RRNPP-mediated QS mechanisms regulate several bacterial processes such as conjugation, sporulation or pathogenicity and is found in many human commensal or pathogenic Gram + bacteria like bacilli, streptococci and enterococci.[240]

Remarkably, RRNPP-like proteins are present outside of the realm of Gram + bacteria, as exemplified by the NIpI proteins from *E. coli*, which contains a lipobox motif that anchors it to the membrane.[241] Furthermore, they can also be found outside bacterial genomes, as illustrated by the recent structure determination of a regulator of phage lysis-lysogeny, AimR.[242]

The effector domains of RRNPP can be classified into three groups: ***i)*** The effector domains of the DNA-binding RRNPP subclass of proteins are helix-turn-helix (HTH), which are motifs that negatively regulate protein expression by binding to DNA, like PrgX form *Enterococcus faecalis*. ***ii)*** Proteins that contain an N-terminal domain with phosphatase activity, which will form a phosphorylation cascade. Examples of these are *B. subtilis* RapA, RapB, RapE, RapH and RapJ. ***iii)*** Proteins that block the action of their cognate effector protein by direct interaction, causing modulation of expression that will initiate a differentiation pathway. This group includes RapC, RapF, RapG, RapH and RapK. RapH is at both second and third groups as it is able to have both activities. Another example of a protein that belongs to more than one group is NprR, as it has a phosphatase and a HTH domain.

One *B. subtilis* RRNPP member is $\text{Rap}_{\text{pLS20}}$. *B. subtilis* encodes 11 Raps on its chromosome[243,244], and at least five plasmid-encoded variants.[245] $\text{Rap}_{\text{pLS20}}$ is the Rap protein encoded by Gram + *B. subtilis* conjugative plasmid pLS20.[181,180] The activity of $\text{Rap}_{\text{pLS20}}$ is regulated by a signaling peptide, $\text{Phr}_{\text{pLS20}}$. This 44 amino acid long pre-peptide is encoded by the small *phr_{pLS20}* gene, which is located immediately downstream of *rap_{pLS20}*, pLS20 gene *26* and *25*, respectively. When secreted, $\text{Phr}_{\text{pLS20}}$ is processed by a second proteolytic cleavage in the extracellular part, which results in the generation of the functional peptide, $\text{Phr*}_{\text{pLS20}}$. The processed peptide is then reimported in an opp-dependent manner. $\text{Phr*}_{\text{pLS20}}$ corresponds to the five C-terminal residues of $\text{Phr}_{\text{pLS20}}$ (QKGMY). The functional peptide is then imported and inactivates $\text{Rap}_{\text{pLS20}}$ by binding to it.

The structure of $\text{Rap}_{\text{pLS20}}$ has been solved by our lab[195]. As it can be seen in **FIGURE 21**, it is all α-helical, consisting of 17 antiparallel helices that are connected by short loops. The C-terminal region consists of 14 helices forming 7 bi-helical TPR motifs, which forms the typical solenoid structure associated with TPR folds. $\text{Rap}_{\text{pLS20}}$ forms dimers in the packaging of the crystal structure in a similar way of *B. subtilis* RapH.[243]

The boundary between the N-terminal effector domain and the C-terminal TPR domain is marked by the loop between H3 and H4. This loop, which consists of 13 residues, is the longest one in the structure and shows poor electron density, indicating a high degree of flexibility.

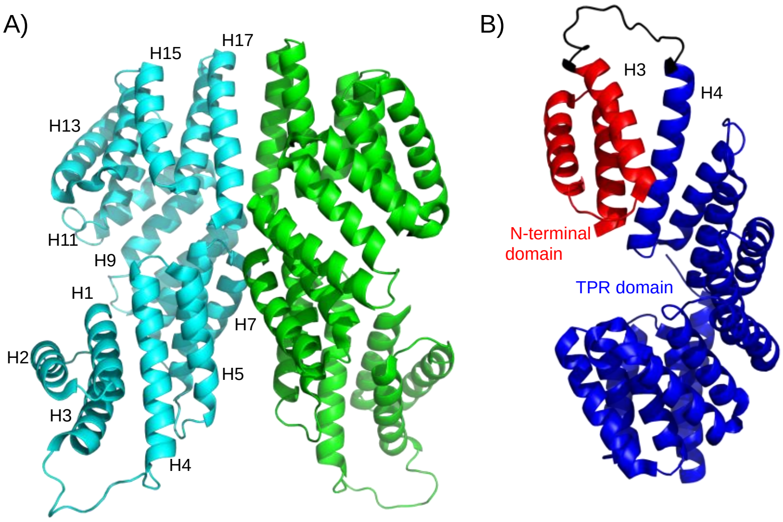


FIGURE 21: **Cartoon representations of the crystallographic structure of $\text{Rap}_{\text{pLS20}}$. A)** The dimeric structure of $\text{Rap}_{\text{pLS20}}$ is shown. **B)** $\text{Rap}_{\text{pLS20}}$ monomer representation distinguishing different domains by colors. N-terminal domain and TPR domains are represented in red and blue, respectively. Loop between H3 and H4 helices is colored in black as it is the border between both domains.

Although $\text{Rap}_{\text{pLS20}}$ apo is similar to most RRNPP proteins, it does have some structural differences with some of the members, occurring in the N-terminal domain. In fact, a number of non-related TPR proteins were found to have higher structural similarity than those of RRNPP proteins, which indicates that these proteins have a common ancestor. The evolutionary link between these proteins lies in the peptide binding function of the TPR domain. For instance, human G-protein-signaling modulator 2, also called LGN for its 10 Leucine-Glycine-Asparagine repeats, has a TPR domain that is structurally similar to that of the Raps. LGN has an important role in mitotic spindle orientation of eukaryotic cell division, and to do so, it binds polypeptide segments of interaction partners.[246] LGN binds peptides that are much longer than the processed signaling molecules in bacterial QS. RRNPP proteins are unable to bind long peptides due to the peptide binding site at the C-terminal forming a cul-de-sac in which only small peptides fit. Another difference between RRNPP and LGN is the N-terminal effector domain, which is not present in LGN.

The structure of peptide bound $\mathrm{Rap}_{\mathrm{pLS20}}$ was also obtained and it is shown in **FIGURE 22**.[195] The approximate location and orientation of the peptide in the TPR domain of $\mathrm{Rap}_{\mathrm{pLS20}}$ is preserved compared with other members of RRNPP family. The peptide is orientated along the solenoid axis of the TPR domain. $\mathrm{Phr*}_{\mathrm{pLS20}}$ binding does not result in any notorious changes in the overall $\mathrm{Rap}_{\mathrm{pLS20}}$ conformation, but it does induce changes on several helices in TPR domain. More concretely, the N-terminal effector domain (NTD) moves outward, away from the solenoid axis of the TPR. The extent of this movement is distinct for the four chains of the $\mathrm{Phr*}_{\mathrm{pLS20}}$-bound $\mathrm{Rap}_{\mathrm{pLS20}}$ and ranges from 2.1Å to 3.5Å. This ample range of movement shows that the TPR domain retains a high level of flexibility in the presence of the peptide.

$\mathrm{Phr*}_{\mathrm{pLS20}}$ promotes a change in the oligomerization state of $\mathrm{Rap}_{\mathrm{pLS20}}$ in a dose-dependent manner, which results in the formation of $\mathrm{Rap}_{\mathrm{pLS20}}$ tetramers. Tetramerization can also be explained by analyzing the crystal structures of both $\mathrm{Rap}_{\mathrm{pLS20}}$ apo and $\mathrm{Rap}_{\mathrm{pLS20}}$ bound to $\mathrm{Phr*}_{\mathrm{pLS20}}$, as an additional interface that results in the tetrameric form of $\mathrm{Rap}_{\mathrm{pLS20}}$ can be found. This was named foot-2-foot interaction and it involves many interactions between the NTDs. Although residues important for this interaction are extremely conserved in some Rap proteins, within the 11 Raps of *B. subtilis* the conservation of the residues is smaller. This suggests that peptide-induced tetramerization is frequent among closely related homologs of $\mathrm{Rap}_{\mathrm{pLS20}}$, but may not occur in the genomic *B. subtilis* Raps characterized so far.[195]

Previous to this finding, even if large conformational changes had been observed in peptide binding for some Raps,[247,248] no related changes in oligomerization states were reported in most Raps. Notable exceptions are PlcR[249] and PrgX[250,251] even if the effect of the peptide binding is different. PlcR tetramerizes upon peptide binding, while two distinct peptides compete for PrgX binding that have stabilizing and destabilizing effects on tetramerization.

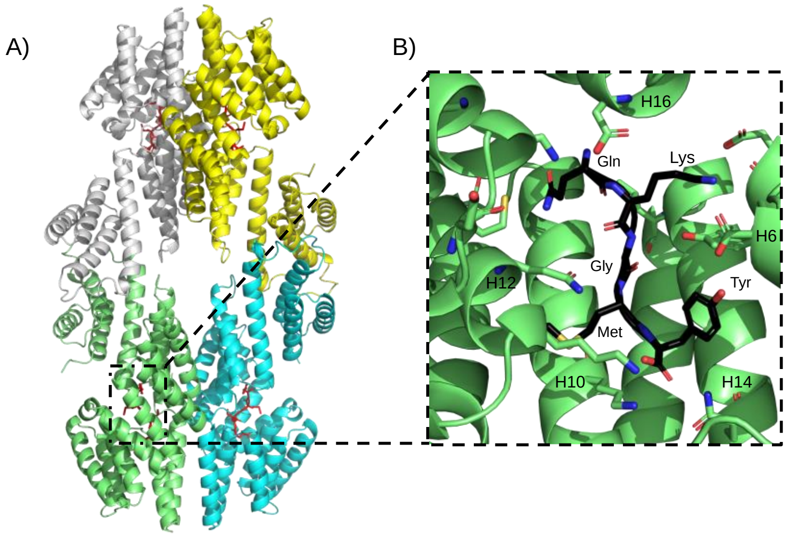


FIGURE 22: **Phr*$_{pLS20}$-induced Rap$_{pLS20}$ tetramerization. A)** Cartoon representation of the Rap$_{pLS20}$ dimer of dimers, colored by chains and displaying the peptide in red. **B)** Stick representation of the close-up view of Phr*$_{pLS20}$ forming contacts with Rap$_{pLS20}$ residues.

Tetramerization has not been observed in any RRNPP members except for NprR and PrgX. In case of PrgX, tetramerization has been suggested to happen via the C-terminal and it is destabilized upon binding to cCf10 peptide. Regarding NprR, NTDs are important for the formation of tetramers and the tetramerization is induced by the peptide, similar to what occurs with Rap$_{pLS20}$.[252] Nonetheless, no foot-2-foot interactions are found in the peptide-bound structure of NprR. The fact that the foot-2-foot tetramerization is only found in Rap$_{pLS20}$ may be due to differences in the NTDs.

In this chapter, we have analyzed by SEC the binding between Rap$_{pLS20}$ and Rco$_{pLS20}$ at different stoichiometries, as well as the effect that Phr*$_{pLS20}$ has in the complex formation. These results are in full agreement with results obtained using different techniques (Small Angle X-ray scattering (SAXS) and analytical ultracentrifugation (AUC)).[253] Furthermore, we have investigated the possibility of cross-regulation by Rap$_{pLS20}$ in other systems that regulate diverse regulatory roles within the cell. To explore the peptide-binding profile of Rap$_{pLS20}$, which is highly correlated with the possibility of cross-regulation, we designed peptides with slightly altered sequences compared to Phr*$_{pLS20}$. These mutations were guided by the peptide Phr*F, cognate peptide of RapF, which has the sequence of QRGMI. We have observed a change in the oligomerization state of Rap$_{pLS20}$ by the addition of some mutated peptides that make Phr*$_{pLS20}$ more similar to Phr*F, but not with the addition of Phr*F itself.

C 2.3. MATERIALS AND METHODS

C 2.3.1. Analytical SEC on $\text{Rap}_{\text{pLS20}}$ and $\text{Rco}_{\text{pLS20}}$ mixtures and the effect of $\text{Phr*}_{\text{pLS20}}$

To determine the elution volumes of the separated proteins, 25µg of $\text{Rap}_{\text{pLS20}}$ (0.56nmol or 22.5µM) and $\text{Rco}_{\text{pLS20}}$ (1.2nmol or 49.4µM) were injected on a Superdex 200 column previously equilibrated in 500mM NaCl, 20mM TRIS pH8, and eluted at a flow rate of 0.2ml min^{-1}. For $\text{Rap}_{\text{pLS20}}$-$\text{Rco}_{\text{pLS20}}$ complex binding stoichiometry tests, molar ratios of 2:1, 1:1, 1:2 and 1:4, respectively, were prepared and incubated 30' on ice before injection. In order to study the effect of the $\text{Phr*}_{\text{pLS20}}$ peptide on $\text{Rap}_{\text{pLS20}}$-$\text{Rco}_{\text{pLS20}}$ complex formation, a 5:1 $\text{Phr*}_{\text{pLS20}}$:$\text{Rap}_{\text{pLS20}}$ stoichiometry was used. 25µl were injected in all cases. All proteins concentrations were in the 0.75mg to 1.25mg ml^{-1} range. The elute was monitored using UV-Vis spectroscopy at 280nm and 260nm. To estimate the M_W of the homo- and heterocomplexes, a calibration of the column was performed as described in §2.3 "Analytical Size Exclusion Chromatography". The derived relation between the V_{el} and M_W was $V_{el} = -0.6815 \cdot \log(M_W) + 5.1906$, with an $R^2 = 0.933$.

C 2.3.2. Analytical SEC on $\text{Rap}_{\text{pLS20}}$ with $\text{Phr*}_{\text{pLS20}}$ and variant peptides

25µg (0,56nmol or 22.5 µM) of $\text{Rap}_{\text{pLS20}}$ was injected in a Superdex 200 column previously equilibrated in 250mM NaCl, 20mM TRIS pH8, 10mM MgCl2, 1mM EDTA, %1 glycerol, 1mM beta-mercaptoetanol. In addition, mixtures of peptides and $\text{Rap}_{\text{pLS20}}$ at 5:1 stoichiometry were injected. Peptides were synthesized *de novo* by an external proteomics service in CBMSO. In case of $\text{Phr*}_{\text{pLS20}}$ at high concentrations, a 10:1 stoichiometry was used. Mixtures were prepared and incubated 30' on ice prior to injection. Elution was done at a flow rate of 0.2ml min^{-1}. The elute was monitored using UV-Vis absorption at 280nm and 260nm. To estimate the M_W, a calibration of the column was performed as described in §2.3 "Analytical Size Exclusion Chromatography". The derived relation between the V_{el} and M_W was $V_{el} = -0.6815 \cdot \log(M_W) + 5.1906$, with an $R^2 = 0.933$.

C 2.4. RESULTS

C 2.4.1. $\mathrm{Rap}_{\mathrm{pLS20}}$-$\mathrm{Rco}_{\mathrm{pLS20}}$ complex formation by analytical SEC

Although we now have structural information about $\mathrm{Rap}_{\mathrm{pLS20}}$ and $\mathrm{Rco}_{\mathrm{pLS20}}$, not much is known about how the two interact to drive regulation of the pLS20 conjugative plasmid. To start to understand this, we analyzed the oligomerization behavior of $\mathrm{Rap}_{\mathrm{pLS20}}$-$\mathrm{Rco}_{\mathrm{pLS20}}$ using analytical SEC. First, $\mathrm{Rap}_{\mathrm{pLS20}}$ and $\mathrm{Rco}_{\mathrm{pLS20}}$ alone were tested. As shown in **FIGURE 23**, $\mathrm{Rap}_{\mathrm{pLS20}}$ and $\mathrm{Rco}_{\mathrm{pLS20}}$ show a very similar V_{el}: $\mathrm{Rap}_{\mathrm{pLS20}}$ elutes at 1.78ml while $\mathrm{Rco}_{\mathrm{pLS20}}$ elutes at 1.80ml. Since SEC allows the determination of the M_W of the eluted peaks, the oligomerization state in solution can be determined from the M_W of the proteins. Under the conditions we tested, $\mathrm{Rap}_{\mathrm{pLS20}}$ behaved as a dimer in solution whereas $\mathrm{Rco}_{\mathrm{pLS20}}$ behaved as a tetramer.

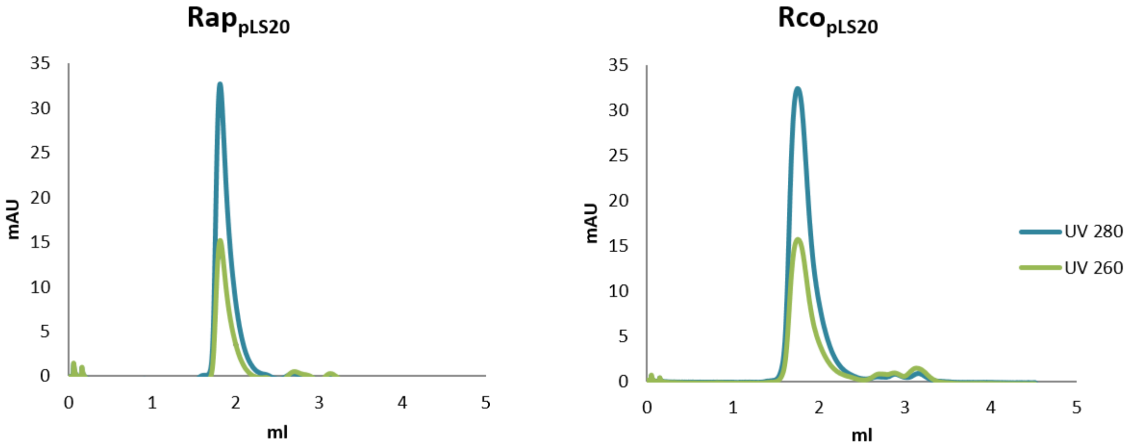


FIGURE 23: **Separation of $\mathrm{Rap}_{\mathrm{pLS20}}$ and $\mathrm{Rco}_{\mathrm{pLS20}}$ by SEC**. Absorbance at 280nm and 260 nm are shown.

The dimerization state of the apo form of $\mathrm{Rap}_{\mathrm{pLS20}}$ is observed for RRNPPs that bind DNA such as PlcR and PrgX, but is uncommon for Raps. Most Rap proteins have been reported to be monomers in solution, including RapF, RapH and RapK[247], and RapJ[248]. Nonetheless, RapH has also been reported to form dimers.[243] Some Rap proteins have been demonstrated to form tetramers as well. The oligomerization states may depend on the presence of the cognate peptide. RRNPP proteins' diverse functions are also revealed by the differences in the oligomerization state and by the effect that the peptide has on oligomerization. It seems that the aggregation behavior of the RRNPPs does not depend directly on the functionality of the N-terminal domain and that there is no common theme in how the oligomerization state of the different RRNPP family members relates to the mechanism.

As mentioned before, $\mathrm{Rap}_{\mathrm{pLS20}}$ is a dimer in solution and seems to form tetramers when bound to $\mathrm{Phr*}_{\mathrm{pLS20}}$. When analyzing the structure, we realize that tetramerization is also reflected in the crystal structure since there is an additional interface resulting in a tetrameric configuration of the Rap_{pLS20} protein. This interface is present both in apo and peptide-bound structures.

Rco_{pLS20} oligomerization state has already been studied previously by Ramachandran *et al.*[180] using AUC. Both sedimentation velocity and sedimentation equilibration experiments were performed. The sedimentation coefficient corrected to standard conditions and the calculated average molecular mass both corresponded to a tetrameric form of Rco_{pLS20}. These results are in line with our results obtained by SEC and the structure solved in §Chapter 1: "Rco_{pLS20}: Key regulatory protein in pLS20 conjugation", as it revealed a tetramerization domain.

As explained in the introductory part of the previous chapter §C 1.2: "Introduction", one of the requirements for the generation of the DNA looping is the oligomerization of the transcription factors. Hereby, Rco_{pLS20} local concentration considerably increases in a given specific target DNA sequence, decreasing the likelihood of nonspecific binding. Up to date, some examples of regulation via DNA looping have been described, which include *ara*, *lac*, *gal*, *deo*, *nag* or *ptsG* operons in *E. coli*. In some of these systems, a supplementary protein is needed to form an open complex. These proteins are commonly known as EBPs (Enhancer Binding Proteins). To the best of our knowledge, there is no EBP in pLS20 but we cannot exclude the possibility of an existence of an auxiliary protein working in coordination with Rco_{pLS20}.

Next, we wanted to determine the stoichiometry of the Rap_{pLS20}-Rco_{pLS20} complex. To do this, the SEC elution profiles of mixtures at different stoichiometries of Rap_{pLS20}-Rco_{pLS20} were analyzed: 2:1, 1:1, 1:2 and 1:4, respectively. These results firmly established that Rap_{pLS20} and Rco_{pLS20} interact to form larger complexes, as indicated by the shift in the V_{el} to lower values, in agreement with previously reported results.[254] The effect of Phr^*_{pLS20} in Rap_{pLS20} alone and the complex was also tested. The SEC elution profile of Phr^*_{pLS20} combined with Rap_{pLS20} showed that a higher order complex is formed, which may be compatible with a Rap_{pLS20} tetramer. Addition of Phr^*_{pLS20} to Rap_{pLS20}-Rco_{pLS20} mixtures resulted in disruption of the Rap_{pLS20}-Rco_{pLS20} complex formation since the peak of the complex disappeared. These data are corroborated by SAXS as well.[253] Interestingly, the Rap_{pLS20}+Phr^*_{pLS20} peaks showed a slight shift to lower V_{el} in the elution profile compared to apo Rap_{pLS20}, indicating the Phr^*_{pLS20}-dependent formation of a Rap_{pLS20} complex of higher M_W, consistent with tetramerization. Therefore, it seems clear that Phr^*_{pLS20} restores the repressive action of Rco_{pLS20} by modulating the direct interaction between Rap_{pLS20} and Rco_{pLS20}. Chromatograms representing these results are shown in **FIGURE 24** and summarized in and **TABLE 4**.

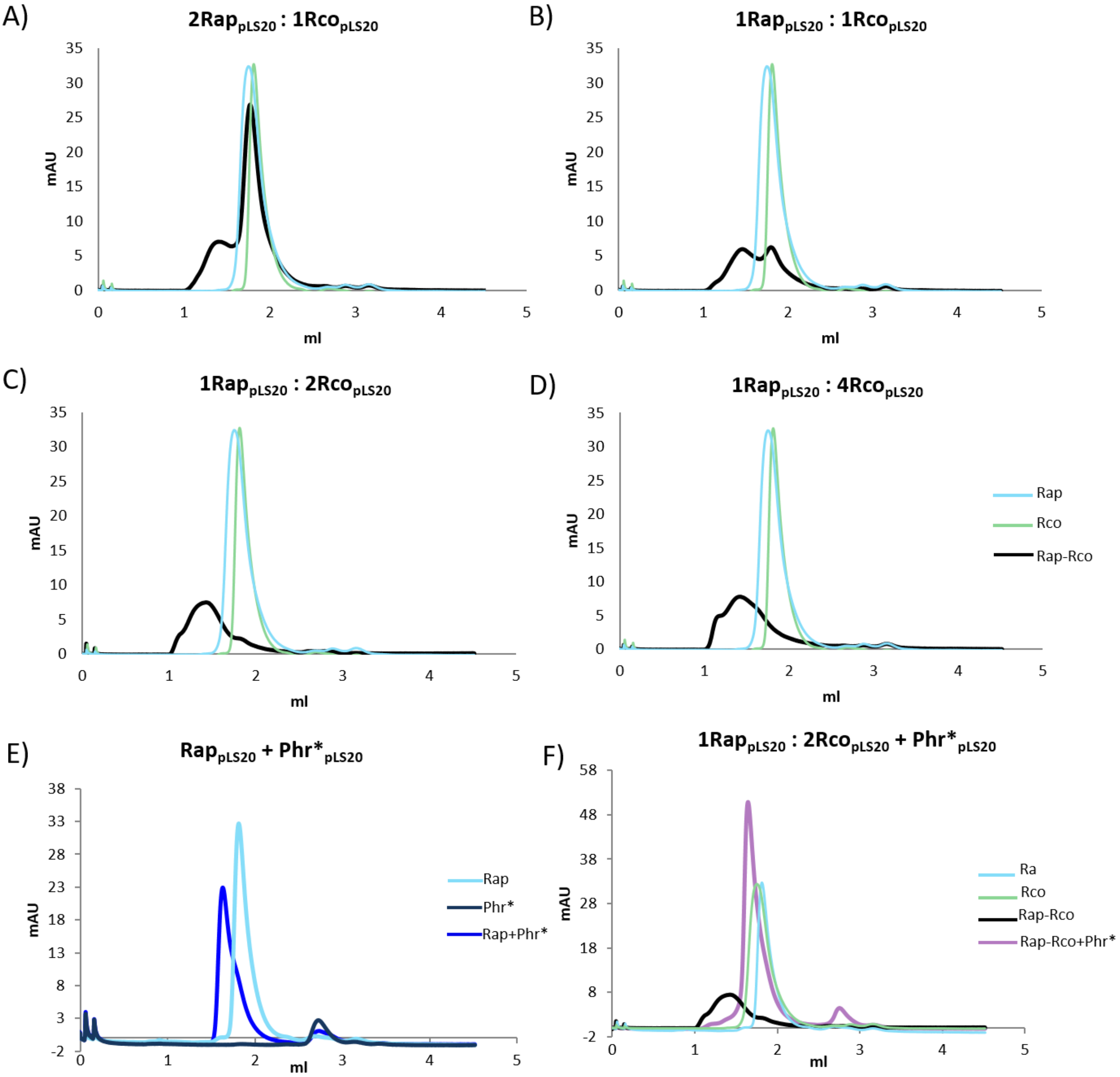


FIGURE 24: **Rap$_{pLS20}$-Rco$_{pLS20}$ complex by SEC.** Different molar ratios of Rap$_{pLS20}$-Rco$_{pLS20}$ were tested: 2:1 (**A**), 1:1 (**B**), 1:2 (**C**) and 1:4 (**D**), respectively. Blue line represents Rap$_{pLS20}$ elution profile, green one corresponds to Rco$_{pLS20}$ and the complex formed between them is shown in black. **E)** Effect of addition of Phr*$_{pLS20}$ to Rap$_{pLS20}$ represented in dark blue. **F)** Effect of the addition of Phr*$_{pLS20}$ to Rap$_{pLS20}$-Rco$_{pLS20}$ complex is represented in light purple. Absorbance at 280nm is shown in all of them.

TABLE 4: **Summary of the results obtained by SEC for Rap$_{pLS20}$-Rco$_{pLS20}$ complex and the addition of Phr*$_{pLS20}$ to it.** The table shows the V$_{el}$ of Rap$_{pLS20}$, Rco$_{pLS20}$, Rap$_{pLS20}$+Phr*$_{pLS20}$, Rap$_{pLS20}$-Rco$_{pLS20}$, Rap$_{pLS20}$-Rco$_{pLS20}$+Phr*$_{pLS20}$, together with an estimation of M$_W$ and predicted oligomerization state.

Protein/complex	V$_{el}$ (ml)	M$_W$ (kDa)	Estimated M$_W$ based on V$_{el}$ (kDa)	Predicted oligomerization state in solution
Rap$_{pLS20}$	1.78	44.43	101	2.27 (Dimer)
Rco$_{pLS20}$	1.80	20.32	94.4	4.65 (Tetramer)
Rap$_{pLS20}$ + Phr*$_{pLS20}$	1.65	45.1	156.68	3.5 (Tetramer)
Rap$_{pLS20}$-Rco$_{pLS20}$	1.47 *	N.A. **	N.A *	N.A. **
Rap$_{pLS20}$-Rco$_{pLS20}$ + Phr*$_{pLS20}$	1.64	N.A. **	162.18	-

*The elution volume is out of the calibration range

** The M$_W$ cannot be calculated since the stoichiometries are unknown

As mentioned before, Rap$_{pLS20}$ forms dimers in solution. However, Phr*$_{pLS20}$ induces a change in its oligomerization state as SEC results demonstrate that an excess of the peptide results in the formation of Rap$_{pLS20}$ tetramers, which suggests Rap$_{pLS20}$ tetramerization may be necessary for Rco$_{pLS20}$ release. This result is in line with results obtained by other techniques like SAXS and AUC. Tetramerization has not been observed in any other RRNPP member up to date except for NprR and PrgX.[253]

The interaction observed between Rap$_{pLS20}$ and Rco$_{pLS20}$ can be understood by studying the Rap$_{pLS20}$ structure. The homotetramerization by the foot-2-foot interactions of the NTDs of Rap$_{pLS20}$ provides an explanation for the inactivation of the Rco$_{pLS20}$ partner. The NTDs are positioned in a manner that enables the interaction with Rco$_{pLS20}$. Upon binding of Phr*$_{pLS20}$ to Rap$_{pL20}$, the NTDs shift outwards, triggering the formation of the homotetramer and changing the interaction surface of the NTDs that are available for interactions with the Rco$_{pLS20}$. Therefore, Rco$_{pLS20}$ is reactivated and able to bind the promoter of the conjugation gene, leading to an inhibition of conjugation. In conclusion, the structural changes induced in Rap$_{pLS20}$ by Phr*$_{pLS20}$ result in the inactivation of the pLS20 conjugation because Rco$_{pLS20}$ is free and binds the repressor.[253]

C 2.4.2. Analytical SEC on Rap$_{pLS20}$ with Phr*$_{pLS20}$ and varying peptides

As it has already been demonstrated, Phr*$_{pLS20}$ binds to Rap$_{pLS20}$ and its structure has been solved (PDB code: 6T46). Phr*$_{pLS20}$ stimulates a change in the oligomerization state of Rap$_{pLS20}$ in a dose-dependent manner, causing Rap$_{pLS20}$ tetramerization.

We wanted to test the hypothesis that Rap$_{pLS20}$ can bind Phr*F, which is the cognate peptide of RapF. We decided to choose this peptide because of sequence similarity with Phr*$_{pLS20}$ (TABLE **5**). RapF initiates the ability for DNA transformation through binding the ComA transcription factor. RapF binds to ComA transcription factor, thereby inhibiting its ability to bind to DNA. RapF and Phr*F act synergistically with RapC and Phr*C[255] in the regulation of ComA.[256]

In order to evaluate the possibility of a cross-regulation between Rap$_{pLS20}$ and RapF systems, we designed some peptide mutants and performed analytical SEC assays. The peptides included Phr*$_{pLS20}$I5Y, in which the tyrosine 5 is mutated to an isoleucine, and Phr*$_{pLS20}$R2K, in which the lysine 2 is mutated to an arginine. By introducing these two mutations together, we obtain Phr*F. Furthermore, we tested the difference in binding when adding Phr*$_{pLS20}$ to Rap$_{pLS20}$ at low and high concentrations to confirm the previous results. The different peptides used are displayed at TABLE **5**.

TABLE 5 **Different peptides tested via analytical SEC.** The mutated residues regarding Phr*$_{pLS20}$ are represented in italics.

Peptides	Sequence
Phr*$_{pLS20}$	QKGMY
Phr*$_{pLS20}$I5Y	QKGM*I*
Phr*$_{pLS20}$R2K	Q*R*GMY
Phr*F	Q*R*GM*I*

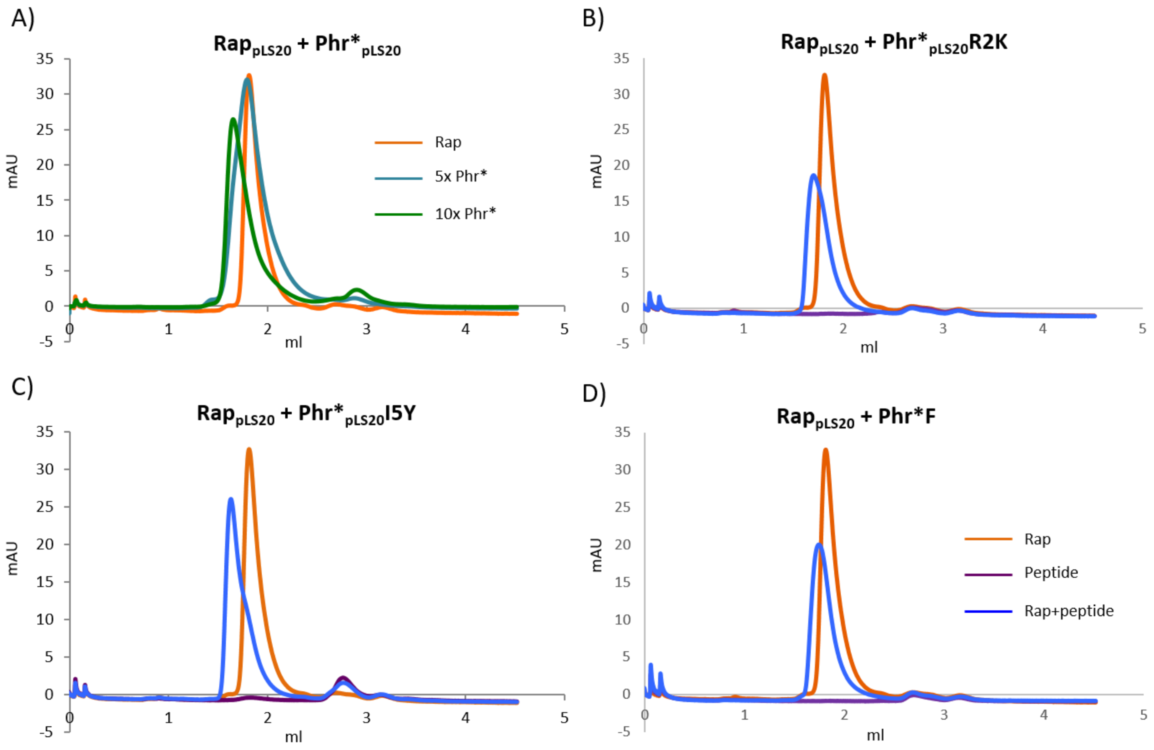


FIGURE 25: **Elution profile of Rap_{pLS20} bound to peptides by SEC. A)** Rap_{pLS20}:$Phr*_{pLS20}$ in an stoichiometry of 1:5 and 1:10. Rap_{pLS20} is represented in orange, 5x$Phr*_{pLS20}$ in light blue and 10x$Phr*_{pLS20}$ in green. Rap_{pLS20} bound to $Phr*_{pLS20}$R2K (**B**), $Phr*_{pLS20}$I5Y (**C**), and Phr*F (**D**). Rap_{pLS20} is shown in orange, peptides in purple and Rap-peptide complex in dark blue. Absorbance at 280nm is shown in all cases.

TABLE 6: **Summary of the results obtained by SEC for Rap_{pLS20} in complex with peptides $Phr*_{pLS20}$ at high and low concentrations, $Phr*_{pLS20}$I5Y, $Phr*_{pLS20}$R2K, Phr*F.** The table shows the V_{el} of Rap_{pLS20} in addition with peptides together with an estimation of M_W and predicted oligomerization state.

Protein + peptide	V_{el} (ml)	Rap M_W (kDa)	Estimated M_W based on V_{el} (kDa)	Predicted oligomerization state in solution
Rap_{pLS20} + low [ ] $Phr*_{pLS20}$	1.76	44.43	108.12	2.43 (Dimer)
Rap_{pLS20} + high [ ] $Phr*_{pLS20}$	1.63	44.43	167.75	3.78 (Tetramer)
Rap_{pLS20} + $Phr*_{pLS20}$R2K	1.70	44.43	132.14	2.97 (Trimer)
Rap_{pLS20} + $Phr*_{pLS20}$I5Y	1.63	44.43	167.75	3.78 (Tetramer)
Rap_{pLS20} + Phr*F	1.74	44.43	115.68	2.60 (Trimer)

As we can see in FIGURE 25 and TABLE 6, Phr*$_{\text{pLS20}}$ induced the dimerization of the dimer of Rap$_{\text{pLS20}}$ at a stoichiometry of 10:1 but not at 5:1. This result agrees with data obtained by other techniques AUX and SAXS, as the tetramerization of Rap$_{\text{pLS20}}$ was also observed in a concentration dependent manner by these techniques. It seems that Rap$_{\text{pLS20}}$ may be in a dimer-tetramer equilibrium that is shifted towards a tetramer form when adding Phr*$_{\text{pLS20}}$, especially at higher concentrations. Interaction of Rap$_{\text{pLS20}}$ and Phr*$_{\text{pLS20}}$ is a crucial requisite for the inhibition of the conjugative genes and, therefore, Rap$_{\text{pLS20}}$ tetramerization may be necessary for conjugation deactivation.

In case of mutants Phr*$_{\text{pLS20}}$R2K and Phr*$_{\text{pLS20}}$I5Y, the V_{el} suggests that Phr*$_{\text{pLS20}}$I5Y induces the tetramerization of Rap$_{\text{pLS20}}$ as it is observed for Phr*$_{\text{pLS20}}$ but it does not seem so clear in the case Phr*$_{\text{pLS20}}$R2K. The V_{el} obtained for Phr*$_{\text{pLS20}}$R2K corresponded to a trimer of Rap$_{\text{pLS20}}$, which may be a mixture between tetramers and dimers. This may be due to the fact of the wide separation range of the column and the associated poor resolution of the peaks. The V_{el} difference between Rap$_{\text{pLS20}}$ dimers and tetramers is relatively small, thus, in case of having a mixture of both oligomerization states, we may not be able to distinguish between the peak that corresponds to dimers and to that of tetramers. Consequently, we may obtain a V_{el} that corresponds to a trimer, as observed for the mutant Phr*$_{\text{pLS20}}$R2K and Phr*F.

It is worth mentioning that the estimation of the M_ws are based on the maxima of the peaks. However, we observed that even if the maximum V_{el} of Phr*$_{\text{pLS20}}$:Rap$_{\text{pLS20}}$ at an stoichiometry of 5:1 corresponded to that of a dimer, the peak was wider towards lower V_{el}s. This may indicate the presence of higher-order oligomers even if the V_{el} maximum value of the convoluted peak corresponds to a dimer.

It did not escape our notice that mutant peptides and Phr*F seemed to induced tetramerization of Rap$_{\text{pLS20}}$ at lower concentrations than its own cognate peptide Phr*$_{\text{pLS20}}$, as 10x Phr*$_{\text{pLS20}}$ was needed to detect the same effect to that obtained by 5x of the other peptides. We propose two possible explanations for this finding. On the one hand, Phr*$_{\text{pLS20}}$ may not have been on its right configuration, meaning the real concentration was lower than 5x. If this were true, the effect observed by 10x would be more comparable to that of the other peptides. On the other hand, the mutant peptides, especially Phr*$_{\text{pLS20}}$I5Y, may have higher affinity to Rap$_{\text{pLS20}}$ and induce its tetramerization at lower concentrations. We believe this is an interesting issue that should be studied in the near future. For example, the binding affinity between mutant peptides and Rap$_{\text{pLS20}}$ could be investigated by fluorescence anisotropy using competitive binding assays as it was done for Phr*$_{\text{pLS20}}$.[195]

We analyzed the contacts of the lysine and tyrosine of the peptide in the structure of its complex with Rap$_{\text{pLS20}}$. Lysine establishes contacts with Y144 and D147, while tyrosine establishes contacts with V141, E145, I182, Q175, K215 and K251. The mutations introduced maintain the nature of interactions, as both lysine and arginine have positively charged side-chains and tyrosine and isoleucine have hydrophobic side chains. However, by introducing the mutations into the structure and subsequently studying the potential contacts arginine and isoleucine may establish with the protein, we realized that less interactions can be expected. Arginine would interact with E145 while isoleucine would interact with Q175, I182 and K215. Contacts of Phr*$_{\text{pLS20}}$

and Phr*F with $\text{Rap}_{\text{pLS20}}$ are represented in **FIGURE 26**. It is important to emphasize that the contacts established by the mutant peptides and Phr*F are simulated and resulted from the simple substitution of the $\text{Phr*}_{\text{pLS20}}$ residues in the $\text{Rap}_{\text{pLS20}}$-$\text{Phr*}_{\text{pLS20}}$ structure. Therefore, we do not have the certainty that these peptides may bind as represented in **FIGURE 26** and we do not know whether it may induce a change in the protein and establish some other contacts. Further investigations are needed.

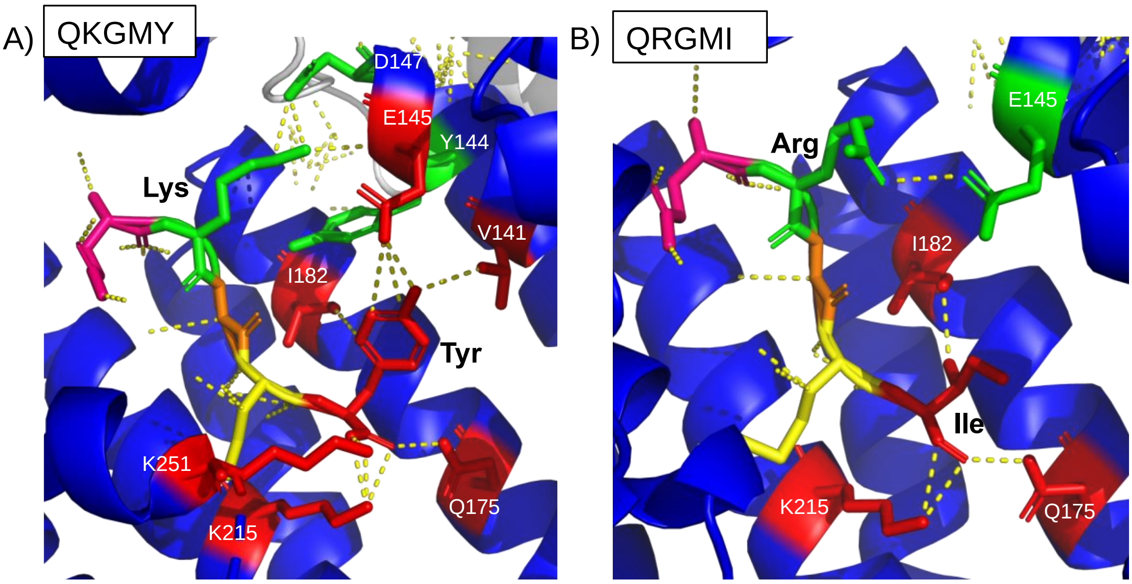


FIGURE 26: **Stick representation of $\text{Phr*}_{\text{pLS20}}$ and Phr*F and its contacts to $\text{Rap}_{\text{pLS20}}$.** $\text{Phr*}_{\text{pLS20}}$ (**A**) and Phr*F (**B**) are displayed showing contacts of lysine and tyrosine (**A**) and arginine and isoleucine (**B**) in green and red, respectively. $\text{Rap}_{\text{pLS20}}$ residues that interact with these are identified by residue name and sequence position. Contacts are displayed as dotted yellow lines.

Obtaining the atomic structure of $\text{Rap}_{\text{pLS20}}$ together with Phr*F-based peptides would confirm that binding occurs and we would be able to visualize their binding mode, thereby, providing a structural explanation for the data obtained by SEC. Hence, we performed crystallization trials with the complexes of $\text{Rap}_{\text{pLS20}}$ with Phr*F-based peptides, using the crystallization condition that resulted in $\text{Rap}_{\text{pLS20}}$+$\text{Phr*}_{\text{pLS20}}$ crystal (PDB code: 6T46).[253] Crystals appeared after 1 week and data to a resolution of about 3Å were obtained. The structure were solved by MR with $\text{Rap}_{\text{pLS20}}$ apo structure (PDB code: 6T3H). However, there was no clear electron density observed at the peptide binding site that corresponded to the Phr*F-based peptides.

Thus, under the conditions we have tested, we observed changes in the oligomerization states of $\text{Rap}_{\text{pLS20}}$ with the addition of peptides $\text{Phr*}_{\text{pLS20}}$R2K, $\text{Phr*}_{\text{pLS20}}$I5Y and Phr*F, suggesting there could be some cross-regulation between both systems, although these results could not be confirmed structurally.

C 2.5. CONCLUSIONS

i. Different Rap$_{\text{pLS20}}$ and Rco$_{\text{pLS20}}$ binding stoichiometries have been tested by SEC and we can confirm the interaction between them.

ii. Phr*$_{\text{pLS20}}$ has been proven to bind to Rap$_{\text{pLS20}}$ and the V$_{\text{el}}$ obtained is compatible with a tetramer, which agrees with data obtained by X-ray crystallography and SAXS.

iii. Also, Phr*$_{\text{pLS20}}$ peptide inhibits the complex formed between Rap$_{\text{pLS20}}$ and Rco$_{\text{pLS20}}$. Consistently, the peak of the Rap$_{\text{pLS20}}$ and Phr*$_{\text{pLS20}}$ complex has a slight shift to lower elution volumes.

iv. Mutant peptides Phr*$_{\text{pLS20}}$R2K and Phr*$_{\text{pLS20}}$I5Y may induce dimerization of the Rap$_{\text{pLS20}}$ dimer as their V$_{\text{el}}$ correspond to a trimer and a tetramer, respectively.

v. In case of Phr*F, the cognate peptide of RapF, a similar effect is observed.

vi. The fact that Phr*F, Phr*$_{\text{pLS20}}$R2K and Phr*$_{\text{pLS20}}$I5Y induce tetramerization of Rap$_{\text{pLS20}}$ at lower concentration than its cognate peptide suggests that their affinity is higher, which should also be investigated by competitive binding assays.

vii. The substitution of the residues K and Y from Phr*$_{\text{pLS20}}$ (QKQMY) is not very drastic and by simple substitutions we have checked contacts with Rap$_{\text{pLS20}}$ can be maintained.

viii. The possibility of cross-regulation between both systems arises and should be further studied.

ix. Attempts to co-crystallize Rap$_{\text{pLS20}}$ and Phr*F-based peptides were not successful. Although diffraction data of relatively good resolution was obtained, we did not see any additional electron density compared to Rap$_{\text{pLS20}}$ apo form.

CHAPTER 3: Reg576, A regulator of establishment

C 3.1. OBJECTIVES

Reg_{576} is a regulatory protein of p576 plasmid of *Bacillus pumilus* system that is an analog of one of the proteins encoded by pLS20 conjugative plasmid. It aids in the establishment of p576 once the plasmid is transferred to the recipient cell. Little is known about how this mentioned establishment occurs, especially in Gram + bacteria. The apo form of the protein has already been solved (PDB code: 6GYG) and DNA binding sequence has been suggested to be the dual heptamer 5´-TTATCCC-3´. Confirming this information would be interesting as it may help finding ways of inactivating Reg_{576} as another possible approach for inhibiting conjugation and therefore, conjugation-mediated spread of ARGs.

These are the specific aims of this chapter:

- Confirming that Reg_{576} binds to the dual heptamer 5´-TTATCCC-3´ and obtaining information about how specific this interaction is by mutating the protein and inducing some changes in the DNA binding region.
- Identifying Reg_{576} residues that may play an essential role in protein stabilization or DNA recognition in order to establish potentially interesting targets for future drug design.
- Solving Reg_{576} structure in complex with its promoter region to visualize how this binding occurs on an atomic resolution, and thereby reveal the specific contacts formed between protein and DNA.
- Obtaining knowledge on the inhibition of gene expression of a newly transferred DNA in the recipient cell, a process which has not been studied much.

C 3.2. INTRODUCTION

The initial steps of the conjugation process involve the formation of a mating pair in which a donor cell recognizes and interacts with a recipient cell. Subsequently, this may trigger a signal that will process the DNA of the conjugative element to generate the T-strand, which will be transferred into the recipient cell. This process needs to be carefully regulated. As stated in the previous chapters, Rap_{pL220} and Rco_{pLS20}, together with $Phr*_{pLS20}$, play a crucial role in the activation/deactivation of the mentioned process. Nevertheless, a successful transfer of the DNA into the recipient cell does not necessarily indicate a successful conjugation event. Once the T-strand arrives into the recipient cell, there are several hurdles that need to be overcome. For example, the ss-DNA must be circularized and converted into ds-DNA, whereas in the case of ICEs, they must be integrated into the recipient cell genome. In addition, for a successful conjugation event, the plasmid should evade the hosts mechanism that bacteria use to defend themselves from foreign DNA.[257]

One of the mechanisms that bacteria use for protecting themselves from foreign DNA is based on the production of methyltransferases (MTase). MTase methylates specific short DNA sequences and thus, shelters these sequences from being digested by the cognate restriction endonuclease (REase) of the restriction modification system. Non-properly methylated DNA that enters the cell is digested by the REase.[258] In response to this, some conjugative elements encode anti-restriction proteins, which are known as Ard (alleviation of restriction of DNA) proteins. Ard proteins impair REase activity. When foreign DNA is properly methylated it is not recognized as foreign and hence, inactivation via REase is relieved. Therefore, anti-restriction genes are necessary for a stable establishment of the conjugative element in the host cell.

Ssb (single-stranded binding) genes also belong to the group of establishment genes. SSB proteins are necessary for coating the single-stranded form of the transferred conjugative element and they bind to ss-DNA without sequence specificity. Plasmid SSB are similar to *E. coli* SSB, which have a role in DNA replication and repair.[259]

PsiB is an example of another establishment gene, which is present in many conjugative plasmids of Gram – origin. During conjugation only one single strand of DNA enters the host cell, hence ss-DNA is a trigger for the recipient cell's SOS response. SOS induction requires activation of RecA, which drives the catalysis of LexA autocleavage. LexA functions as a repressor of SOS genes. PsiB protein binds RecA, inhibiting RecA filament formation on SSB-coated ssDNA. This inhibition is sufficient to prevent the RecA-catalyzed LexA autocleavage required to induce SOS response.[260,261]

For establishment genes to function properly, they must be expressed rapidly upon entry of the conjugative element. Importantly, prolonged expression of these genes may be harmful for the cell. Inactivation of the restriction factors for long periods of time makes the cell vulnerable to entry of any foreign DNA, such as phage DNA. Moreover, extended inhibition of the SOS response may result in the cell not responding correctly to different stimuli like DNA damage. Therefore, there are specific regulatory mechanisms that ensure that establishment genes are expressed rapidly but transiently after the transfer of the conjugative element in the recipient cell.

Most information about the expression of establishment genes comes from studies done on conjugative *E. coli* plasmids, more concretely F and ColIB-P9 plasmids. The *ssb*, *ard* and *psiB* genes are present in enterobacterial plasmids that belong to non-compatible groups.[262,263] There are specific promoters that are active when the DNA is in its single-stranded form[264] that are in charge of controlling these genes so they are expressed rapidly and transiently when DNA enters the recipient cell.[265,266] As a result, ss-DNA entering the cell triggers their expression and ss-DNA conversion to ds-DNA inhibits the expression.

One of the genes regulated in this manner is *ardC$_{576}$*, which belongs to conjugative plasmid p576 of *Bacillus pumilus*. Its encoded protein's sequence has approximately 50% similarity with type C anti-restriction proteins encoded by conjugative plasmids of Gram – bacteria that function as establishment genes of plasmids in recipient cell.[265,267] Therefore, it seems that *ardC$_{576}$* encodes an anti-restriction protein may help in the establishment of p576 in a recipient cell. P576 is closely related to pLS20.[267]

P$_{20c}$, P$_{23C}$, P$_{reg576}$ (P$_{27c}$) and P$_{ardC576}$ are p576 promoters. It was observed that next to all of these promoters there were dual boxes of the heptamer sequence 5'-TTATCCC-3', to which Reg$_{576}$ binds as determined by EMSAs (electrophoretic mobility shift assay) performed by Val-Calvo *et al.*[268] All of these promoters are regulated in a similar way. They are initiated when the plasmid is transferred into the recipient cell and once enough Reg$_{576}$ is produced, these promoters are inhibited again. Moreover, it has been shown that Reg$_{576}$ regulates the expression of its own promoter P$_{27C}$. P$_{27C}$ is a weaker promoter than P$_{20C}$ and P$_{23C}$ and it also differs in terms of configuration. This is probably due to the fact that P$_{20C}$ and P$_{23C}$ contain two Reg$_{576}$ operators while P$_{27C}$ contains only one. P576 promoters are shown in **FIGURE 27**.

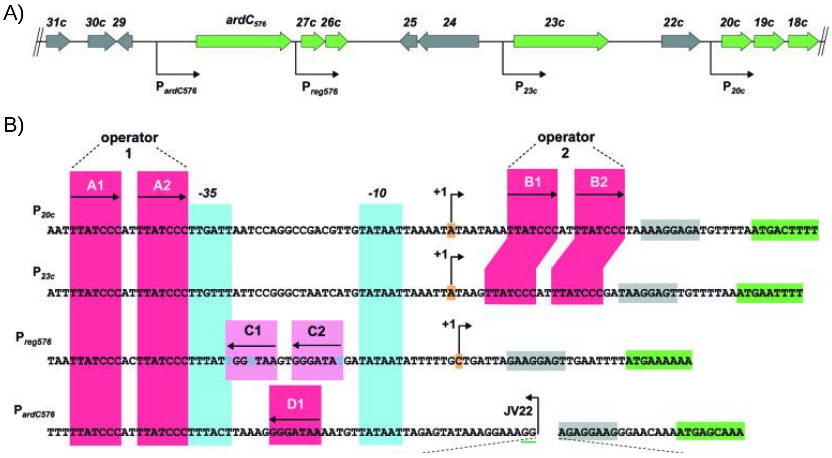

A)
31c
30c
29
ardC576
27c
26c
25
24
23c
22c
20c
19c
18c
PardC576
Preg576
P23c
P20c
B)
operator 1
operator 2
A1
A2
-35
-10
+1
B1
B2
P20c
AATTTATCCCATTTATCCCTTGATTAATCCAGGCCGACGTTGTATAATTAAAATATAATAAATTATCCCATTTATCCCTAAAAGGAGATGTTTTAATGACTTTT
+1
P23c
ATTTTATCCCATTTATCCCTTGTTTATTCCGGGCTAATCATGTATAATTAAATTATAAGTTATCCCATTTATCCCGATAAGGAGTTGTTTTAAATGAATTTT
C1
C2
+1
Preg576
TAATTATCCCACTTATCCCTTTAT GG TAAGTGGGATA GATATAATATTTTTGCTGATTAGAAGGAGTTGAATTTTATGAAAAAA
D1
JV22
PardC576
TTTTTATCCCATTTATCCCTTTACTTAAAGGGGATAAAATGTTATAATTAGAGTATAAAGGAAAGG AGAGGAAGGGAACAAAATGAGCAAA

FIGURE 27: **5'-TTATCCC-3' directed repeats located near the p576 promoters P$_{20c}$, P$_{23c}$, P$_{27c}$ and P$_{ardC576}$. A)** Genetic organization of p576 region. Genes are represented with arrows, green ones corresponding to genes controlled by promoters P$_{20c}$, P$_{23c}$, P$_{27c}$ and P$_{ardC576}$. **B)** Sequences of promoters P$_{20c}$, P$_{23c}$, P$_{27c}$ (P$_{reg576}$) and P$_{ardC576}$. Heptamer sequences 5'-TTATCCC-3' are highlighted in pink. Gray boxes and green boxes represent RBS (ribosome binding site) and coding sequences, respectively. *"Image taken from doi: 10.1093/nar/gky996"*

The structure of Reg$_{576}$ has been determined (PDB code: 6GYG).[268] Reg$_{576}$ is a Ribbon-Helix-Helix (RHH) protein that contains an additional third α-helix (H3). In some other RHH proteins an additional C-terminal helix is also present, such as, Mnt (PDB code: 1MNT) or SSV-RH (PDB code: 4AAI). Nonetheless, the placement of this third helix is different for Reg$_{576}$ when compared to Mnt or SSV-RH.

In prokaryotes, most DNA transcription factors have a Helix-Turn-Helix (HTH) motif that places an α-helix into the DNA major groove to bind to an specific operator.[269] For some time, it was thought that α-helixes were the only motifs prokaryotes used for binding to DNA. However, a different binding model was proposed for Arc and Mnt repressors of bacteriophage P22 based on NMR structure of Arc and some biochemical data, which showed that specificity determinants for the binding were placed within the amino-terminal β-strands of the Arc and Mnt dimers.[270,271,272] In 1992, the structure of MetJ with its operator DNA was solved and it was clear how an antiparallel β-sheet was present at the N-terminal of the dimer and it was located in the DNA major groove to make specific nucleotide base contacts.[273] Shortly after that, in 1994, the structure of Arc-DNA complex was solved, which also revealed a specific DNA recognition by a β-sheet (PDB code: 1PAR).[274] This discovery led to the identification of a new transcription factor family, named the RHH superfamily based on the order of their secondary structure elements (β-strand, α-helix, α-helix).

The superfamily of RHH transcription factors is composed of proteins that are involved in the regulation of several bacterial processes like cell division, control of plasmid copy number, amino acid biosynthesis or lytic acid cycle of bacteriophages.[275] RHH transcription factors have been found in bacteria, archaea and bacteriophages. It has been demonstrated that many auxiliary relaxosome proteins encoded by conjugative plasmids of Gram + bacteria hare also RHH type of proteins.[276], in fact structure-based mutational analyses have shown how important the RHH motif is in *oriT* binding and relaxase recruitment.[277] RHH proteins use a relatively short conserved 3D structural domain (generally formed of 36 residues) to bind DNA in a sequence-specific manner. Most RHH proteins are dimers having a 2-fold symmetry since this 3D structural domain is formed by intertwining of the RHH motif of two monomers referred to as RHH_2. Usually, RHH proteins bind operators composed of two or more binding sites organized as inverted or tandem repeats to which higher order RHH oligomers bind.

The RHH core consists of a dimeric structure in which a double stranded anti-parallel β-sheet is formed by two N-terminal β-strands from the monomers. RHH proteins bind DNA by intercalation of their 2 β-strands into the major groove of ds-DNA, this interaction contributes to affinity and specificity of the binding.[275] Within the N-terminal β-strand, there are some residues that share a structurally equivalent function among different auxiliary factors. These residues have been demonstrated to be necessary for DNA binding, as shown by solved structures of TraM or ArcA bound to DNA.[274,278] Mutation of surface-exposed charged residues impairs binding of PcfF to pCF10 *oriT*.

Regarding Reg_{576}, its operator is composed of two subunits that are arranged in a directed repeated orientation spaced by 2bp. Two Reg_{576} dimers bind to this operator. Reg_{576} is thought to bind with a very high level of cooperativity to its promoters when compared to other RHH members, based on the observation that binding between Reg_{576} and ds-DNA containing a single heptamer results impaired based on EMSA and AUC data. This is not always the case for some other RHH proteins like Arc.[268]

Several drugs have been developed to target relaxase activity but none that target establishment genes. This may be due to the fact that the information about these genes and their regulation in Gram + bacteria is scarce.

In this chapter, we tested the binding between Reg_{576} and various ds-DNAs containing a single copy and a direct repeat (DR) of the heptamer by analytical SEC. We will refer to this heptamer as regRE. Also, a mutation was introduced in a residue that seemed to be important for Reg_{567} functionality. By using this mutant protein for SEC assays, we determined it was not a crucial residue as we could still observe binding to DR of regRE. Once the binding of Reg_{576} was confirmed, we tried to elucidate the Reg_{576} structure in complex with its binding domain by X-ray crystallography. Nevertheless, we did not succeed in this purpose and instead we obtained several additional Reg_{576} structures with different crystalline packages and space groups. Among these, a potentially interesting structure was obtained in terms of binding manner to DNA.

C 3.3. MATERIALS AND METHODS

C 3.3.1. Reg_{576} and ds-DNAs binding assays by analytical SEC

Reg_{576} and Reg_{576}46K>A were purified as explained in §2.2 “Chromatography techniques and protein purification protocol”. 37.5µg (3.44nmol or 137.6 µM) of Reg_{576} were injected both on a Superdex 75 and a Superdex 200 columns previously equilibrated in 250mM NaCl, 20mM TRIS pH8. The elution was done at a flow rate of 0.3ml min^{-1}. Also, ds-DNA binding tests were performed. Complementary strands of ss-DNA, synthesized by Biomers, were combined at a stoichiometry of 1:1. These mixtures were annealed at 95°C and were then allowed to cool down to room temperature overnight. Tested ds-DNAs are listed in TABLE **7**. For each ds-DNA, a mixture of a molar ratio of 4:1 of protein:ds-DNA (3.44nmol:13.76nmol) was prepared and incubated for 30’ on ice before injection. The protein was either native Reg_{576} or Reg_{576}46K>A. 25µl were injected in all cases. The elute was monitored using UV-Vis spectroscopy at 280nm and 260nm. The M_W of the elution peaks was determined using a calibration equation derived as described in §2.3 “Analytical size-exclusion chromatography”.

TABLE 7: **Different ds-DNAs tested via analtyical SEC.** The Reg_{576} binding site is represented in yellow. A single strand is represented although all were tested in a double-stranded form.

ds-DNA	Nucleic acid Sequence
22A	TTAATTATCCCACTTATCCCTT
22B	AATTATCCCACTTATCCCTTTA
24	TTAATTATCCCACTTATCCCTTTA
20	AATTATCCCACTTATCCCTT
9	ATTATCCCA
12	ATCCCACTTATC

C 3.3.2. Reg$_{576}$ Crystallization and Data collection

Reg$_{576}$ was purified as described in §2.2 "Cromatography techniques and Protein purification protocol". A second purification based on its particle size (M_W: 11kDa) was done using ProteoSEC 3-70HR column (Generon) using 500mM, 20mM TRIS pH8 as a buffer. We performed several trials to co-crystallize Reg$_{576}$ with DNA. The ds-DNAs tested were 22A, 22B, 20 and 24 (TABLE **8**), for which binding was confirmed by SEC.

For the crystallization trials, we used commercial crystallization screenings listed in §2.4 "Crystallization techniques", with protein concentrations ranging from 8 to 16mg ml^{-1}, stoichiometries of 2:1 and 4:1 (protein:DNA) and protein buffers containing 500mM NaCl or 250mM NaCl with 20mM Tris-HCl pH8. In addition, crystallizations were performed on apo Reg$_{576}$.

Crystals that correspond to the structure of Reg$_{576}$ described below were obtained by the sitting drop vapor-diffusion method at 18°C, by equilibration of drops of 100nl of protein + 100nl of crystallization buffer (1.1M Sodium malonate dibasic monohydrate, 0.1M Hepes pH7, 0.5% v/v Jeffamine® ED-2003). Protein concentration was of 8mg ml^{-1} being its buffer 250mM NaCl, 20mM Tris-HCl pH8. Crystals were harvested after three months of incubation and they were cryo-cooled by direct transfer from the crystallization drop into liquid nitrogen for X-ray collection.

C 3.3.3. Reg$_{576}$ Diffraction and Data processing

Data collection was performed at ALBA Synchrotron Light Source on the BL13-Xaloc beamline with an X-ray photon energy of 12.66 keV.[196] Data was processed the AutoPROC toolbox (Global Phasing Ltd.[197]), using anisotropic resolution cutoffs.[198] Structures were determined by MR using PHASER[168] as implemented in the CCP4 suite of programs.[280] PDB entry 6GYG[268] was used as a search model.

The model was refined automatically with Phenix.refine[281], REFMAC5 from CCP4 suite[201] and also manually with COOT.[282]. Figures were prepared using PyMOL (The PyMOL Molecular Graphics System version 2.3 Schrödinger, LLC).

C 3.4. RESULTS

C 3.4.1. Identification of the residues involved in the binding between Reg_{576} and DNA by analytical SEC

Analytical SEC assays were performed to confirm the binding of Reg_{576} to the recognition sequence in its operator, which is composed of two heptanucleotide sequences that are arranged in a DR, spaced by 2bp. This dual heptamer is present in all Reg_{576} promoters: P_{20c}, P_{23C}, P_{27c} and $P_{ardC576}$. (**FIGURE 27**) Several ds-DNA sequences were tested, which are represented in TABLE **8**. Most of these sequences contained a DR of regRE (22A, 22B, 24, 20) but one of them contained a single regRE (9). Furthermore, another type of ds-DNA (12) was also assessed: a sequence lacking the full DR of regRE, but having the 2bp that separate the repeats to evaluate the possibility of Reg_{576} binding to the central part of the DR of regRE instead of to the 5'-TTATCCC-3' itself.

We first tested the oligomerization state of Reg_{576} using SEC. For that purpose, Reg_{576} alone was injected into two different analytical columns: Superdex 75 and Superdex 200, which have separation ranges of 3-70kDa and 6-600kDa, respectively. We obtained a V_{el} of 1.54ml in the Superdex 75 column and a V_{el} of 2.18ml in the Superdex 200 column. Based on our in-house made calibration curves, the protein would have a M_W of 23.7kDa regarding the Superdex 75 column and a M_W of 26.1kDa based on Superdex 200 column. Reg_{576} having approximately 11kDa, it seems that our protein would be dimer in solution under the conditions we tested.

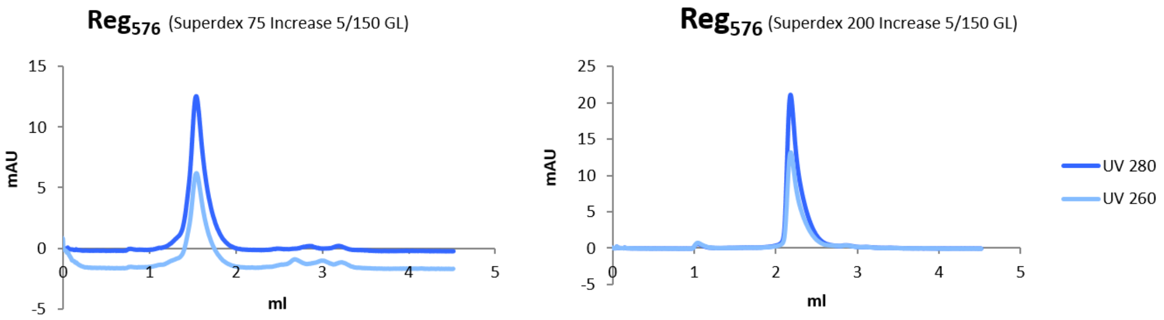


FIGURE 28: **Separation of Reg_{576} by two different analytical SEC columns.** The left graph was done in a Superdex 75 column whereas the right graph corresponds to a Superdex 200 column. Absorbance at 280nm and 260nm are represented in dark blue and light blue, respectively.

The oligomerization state of Reg_{576} *in vitro* has already been studied by Val-Calvo *et al.*[268] by analytical ultracentrifugation sedimentation velocity (SV), sedimentation equilibrium (SE) and dynamic light scattering (DLS). In the SV analysis, Reg_{576} was observed as a single species whose corrected experimental sedimentation coefficient was consistent with a dimer. Regarding DLS assays, a M_W of 22kDa was obtained based on the translational diffusion coefficient (D) and the S-value obtained from SV analysis. A molar mass of 21.8kDa was calculated using SE. All of the results obtained by parallel and complementary techniques, including our SEC data, show that Reg_{576} forms dimers in solution. Importantly, most RHH proteins are dimers having a 2-fold symmetry since the quaternary structure of these domains is formed by intertwining of the RHH motif of two monomers referred to as RHH_2. Indeed, most RHH proteins recognize their binding sites by forming dimers of dimers. This results in an increased specificity of DNA binding.

Next, binding of Reg_{576} to dual heptamer containing ds-DNAs was tested (FIGURE 29). As mentioned before, ds-DNAs 22A, 22B, 24 and 20 contain two direct repeats of the 5'-TTATCCC-3', recognition element, as shown in TABLE **7.** By this first SEC tests we confirmed that Reg_{576} is binding to the dual heptamer 5'-TTATCCC-3', as we see a shift in the V_{el} of the complex formed by Reg_{576} + ds-DNA towards to higher M_W compared to that of the protein and ds-DNA alone.

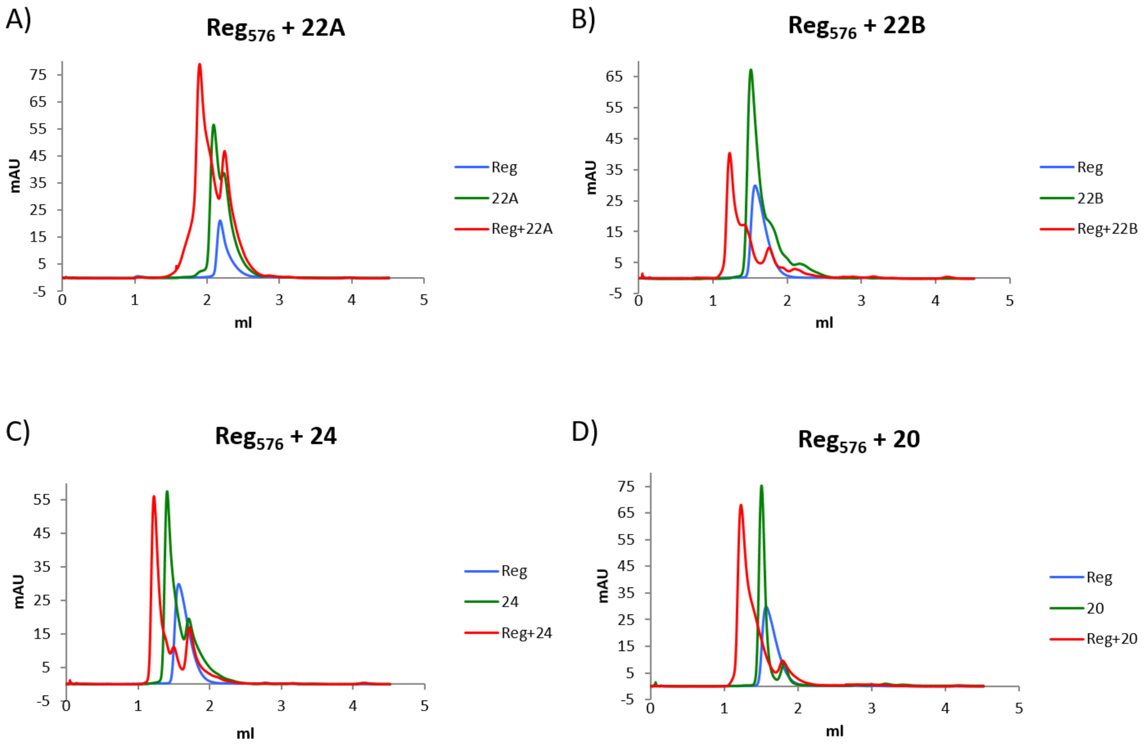


FIGURE 29: **Reg_{576} binding to dual regRE containing ds-DNAs by SEC.**The elution profiles of mixture or Reg_{576} with ds-DNAs 22A (**A**), 22B (**B**), 24 (**C**) and 20 (**D**) are shown. Reg_{576} is represented in blue, ds-DNAs in green and complex between Reg_{576} and ds-DNAs in red. Absorbance at 280nm is shown in all the graphs. (**A**) Reg_{576}+22A was performed using a Superdex 200 column, whereas the binding assays with the rest of ds-DNAs 22B (**B**), 24 (**C**) and 20 (**D**) were performed in a Superdex 75 column.

After having confirmed that Reg_{576} binds to all tested ds-DNAs that contain the DR of regRE, we decided to further test binding of Reg_{576} to the ds-DNA containing a single copy of the regRE (9) and the ds-DNA containing the central part of the DR of regRE (12). Using these ds-DNAs, we were able to test whether Reg_{576} has affinity for a ds-DNA containing a single copy of regRE, and whether it is able to bind a ds-DNA containing the internal sequence of the DR of the regRE sequences, to discard the specificity is conferred by the middle region of the DR instead of by the DR itself.

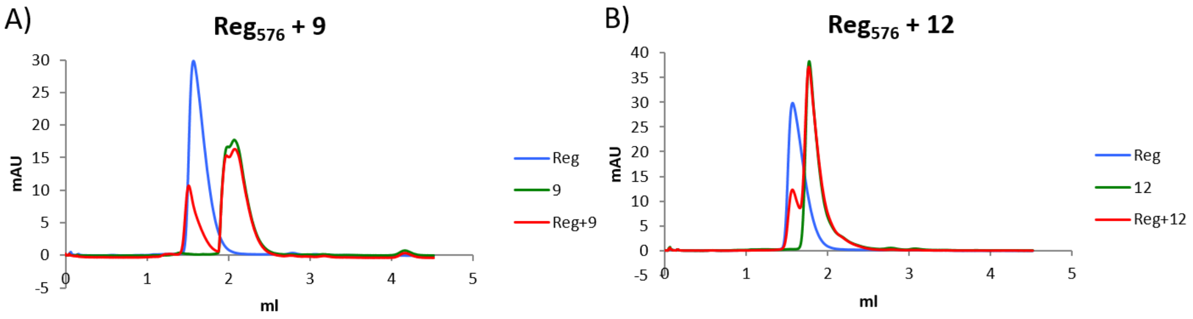


FIGURE 30: **Reg_{576} binding to ds-DNAs that do not contain the dual heptamers by SEC. A)** ds-DNA 9 (**A**) contains only a single heptamer and ds-DNA 12 (**B**) contains the middle region of the DR. Reg_{576} is represented in blue, ds-DNAs in green and complex between Reg_{576} and ds-DNAs in red. Absorbance at 280nm is shown in all graphs.

As shown in **FIGURE 30**, Reg_{576} does not bind to ds-DNA containing a single regRE or to a ds-DNA containing only the middle region of the longer ds-DNAs 22A, 22B, 24 and 20. This latter observation, combined with previous results showing that Reg_{576} binds tightly to direct repeat sequences of regRE also containing the fragment of ds-DNA 12, shows that the affinity of Reg_{576} is specific for this dual 5'-TTATCCC-3' sequence. This can be explained by the fact that Reg_{576} may bind to its operator with a very high level of cooperativity, as confirmed by previous studies.[268] In fact, its binding affinity seems to be higher compared to other RHH proteins. For instance, RHH Arc protein of bacteriophage P22 binding to only one single Arc dimer to a DNA fragment containing a single subsite was observed by EMSA.[283] On the contrary, no Reg_{576} binding was detected to ds-DNA that contained an individual subsite by EMSA[268], consistent with our SEC results regarding binding to ds-DNA 9.

As mentioned in the introductory part of the chapter, Reg_{576} contains the RHH fold, but has an additional C-terminal α-helix. Mnt and SSV-RH also contain a third α-helix but they are differently disposed. In Reg_{576}, helix H3 contacts the outward flank of α-helix1 of the opposite monomer (H1'). In Mnt, the third helix folds back on the H2', but on the opposite side compared to Reg_{576}, interacting with H1'. In SSV-RH, H3 folds back over the external flank of H2'. The presence of an additional α-helix and its particular position with respect to the RHH core distinguishes Reg_{576} from other RHH proteins. In fact, it is the only known protein with this type of geometry, with the exception of a protein from *Nitrosomonas europaea*, whose function is unknown (PDB code: 1ZX3)[268] (**FIGURE 31**).

Importantly, by doing structural alignments we have realized that the lysine 46 is conserved in the RHH proteins with an additional α-helix (**FIGURE 31**). Lysine 46 is located at the loop that connects helices H2 and H3. Residues from K46 to E49 constitute this loop. H3 contacts the H1 of the other monomer. This led us to think that this lysine residue may have an important role. In RHH proteins, DNA binding occurs through intercalation of their β-sheets within the RHH domain into the major groove of ds-DNA, therefore, this lysine residue is not involved in the DNA recognition but it may be crucial for the cooperativity of the complex formed between Reg$_{576}$ dimers. Based on this, we decided to introduce a mutation of the lysine at position 46 by an alanine (Reg$_{576}$46K>A).

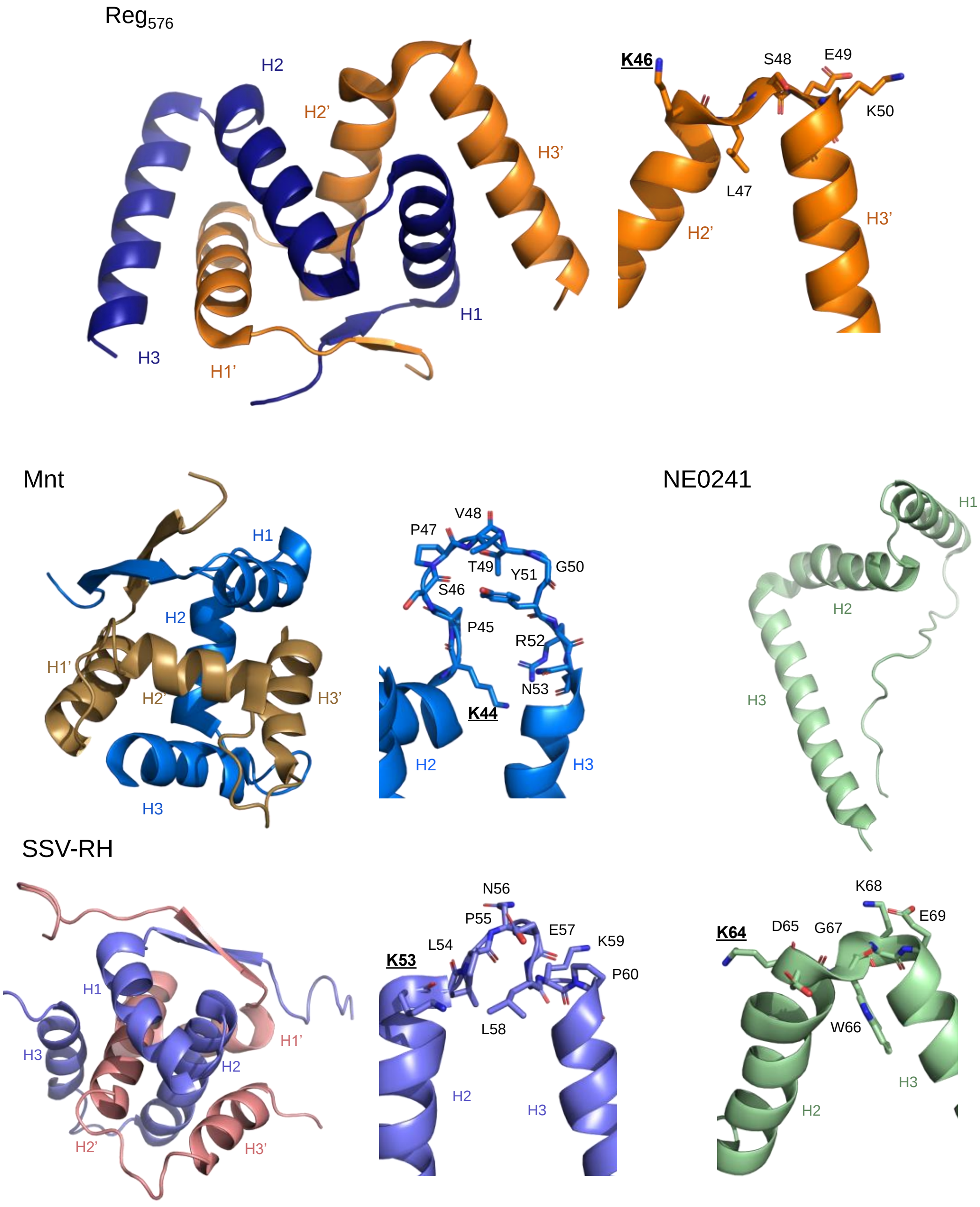


FIGURE 31: **Cartoon and stick representation of Reg$_{576}$, Mnt, SSV-RH and hypothetical protein NE0241 from *N. europeae.*** Overall structure of Reg$_{576}$ (PDB code: 6GYG), Mnt (PDB code: 1MNT), SSV-RH (PDB code: 4AAI) and NE0241 (PDB code: 1ZX3) are shown together with a stick representation of the loop between H2 and H3. Conserved lysine is highlighted in bold.

We next performed an analytical SEC using the mutant protein Reg_{576}46K>A and ds-DNA containing the DR of regRE (22B). The result shows that binding between Reg_{576} and its binding site does occur when this specific residue is mutated. Remarkably, the V_{el} of the complex formed by the native and mutant protein with ds-DNA 22B is similar, suggesting that they may bind DNA in the same way.

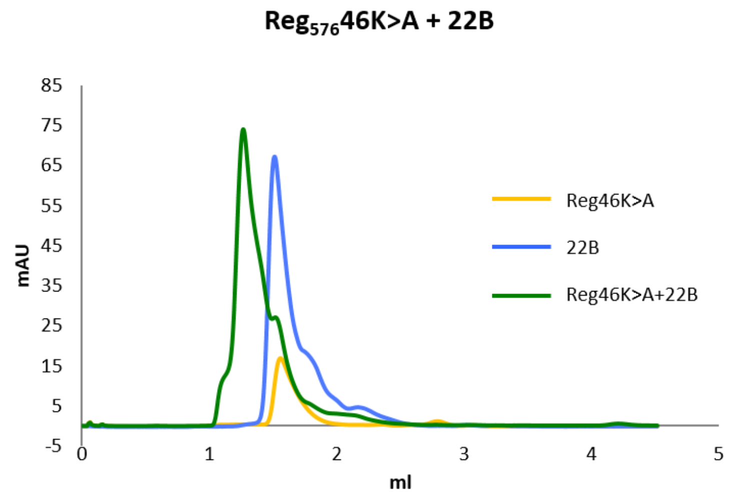


FIGURE 32: **Reg_{576}46K>A binding to dual regRE containing ds-DNA by SEC.** Mutant protein is represented in yellow, ds-DNA in blue and complex between protein and ds-DNA in green. Absorbance at 280nm is shown.

Therefore, K46 does not seem to be crucial for binding to its operator region, as when this residue is mutated, binding to regRE is not impaired. K46 is not the only residue involved in sustaining the complex formed between Reg_{576} dimers. It would be interesting to study the effect of removing the full H3 in the stabilization of the dimer and DNA binding. Our hypothesis is that H3 is in charge of sustaining the complex formed by Reg_{576} dimer and its operator.

To elucidate the structural basis of complex formation between Reg_{576} and DNAs containing a DR of regRE, the structure of the complex should be obtained. Although the binding mode is similar to other RHH protein such as Arc (PDB code: 1PAR), CopG (PDB code: 1EA4) and AmrZ (PDB code: 3QOQ), which all bind to DNA repeats that are separated by approximately one turn of the DNA double helix, it seems like Reg_{576} may involve more interactions between the two dimers.[268] Furthermore, by revealing the contacts formed between Reg_{576} and DNA, we would fully understand the reason for the observed increased binding cooperativity.

C 3.4.2. Reg_{576} tetrameric structure

A lot of experiments were performed to obtain a co-crystal of Reg_{576} with ds-DNAs obtaining the DR of regRE. Unfortunately, we did not succeed in this goal. Nevertheless, and even if Reg_{576} apo structure had already been solved (PDB code: 6GYG), an interesting novel structure of Reg_{576} was obtained in a different space group and crystalline package.

The statistics of the processing of the X-ray diffraction data, the refinement of the final structure and its validation are summarized in TABLE **8**.

TABLE 8: **X-ray processing and refinement statistics for Reg_{576}.**

Data processing statistics	**Reg_{576}**
Space group	$P\ 3_2\ 1\ 2$
Unit-cell parameters (Å)	a=81.947 b=81.947 c=79.092
Unit-cell angles (°)	α=90 β=90 γ=120
Resolution range (Å)[a]	2.113-70.968
No. of unique reflections	17601
Completeness (%)	100
Redundancy	19.5
Mean I/σ(I)	14.6
R_{meas} (%)[b]	0.098
Refinement statistics	
R_{work}[c] (%)	25.49
R_{free}[d] (%)	31.84
Ramachandran	
Favored (%)	94.83
Disallowed (%)	0.00
R.M.S.D.	
Bond lengths (Å)	0.009
Bond angles (°)	1.07
Chirality	
Mean B value (Å^2)	82.6

[a] Numbers in parentheses represent values in the highest resolution shell.

[b] Rmeas=$\sum hkl\ [N/N-1]^{1/2} \sum_i |Ii(hkl) - \langle I(hkl)\rangle| / \sum_{hkl} \sum_i Ii(hkl)$ where N is the multiplicity of a given reflection, Ii(hkl) is the integrated intensity of a given reflection, and <I(hkl)> is the mean intensity of multiple corresponding symmetry-related reflections.

[c] Rwork=$\sum ||Fobs| - |Fcalc|| / \sum |Fobs|$, where |Fobs| and |Fcalc| are the observed and calculated structure factor amplitudes, respectively.

[d] R_{free} is the same as R_{work} but calculated with a 5% subset of all reflections that was never used in refinement.

We cannot give an explanation as to why the Reg_{576} complex with ds-DNA resisted crystallization despite the large number of trials performed (described at §C 3.3.2 "Reg_{576} crystallization and Data Collection") specially considering that the binding was confirmed by SEC.

Initially, the protein-DNA complex was in a buffer containing 500mM NaCl, 20mM TRIS pH8. We suspected salt might be impairing the binding between the protein and the DNA, since it is a major factor as a high salt concentration reduces binding through electrostatic interactions. Based on this, we performed further trials with a buffer that contained 250mM NaCl, 20mM Tris-HCl pH8. However, we were not able to decrease the concentration more than that due to protein stability problems. Regarding stoichiometry, we tested 4:1 (protein:DNA) firstly, based on the analytical SEC assays in which we did observe binding under these conditions.

Furthermore, it is still more striking not having been able to fulfill our goal of obtaining structural information about the binding between Reg_{576} and regRE given the measured high cooperativity of the complex.[268]

It is worth pointing out that when performing the crystallization trials with DNA, the same conditions were applied to the protein in its apo form as a negative control. Many crystals in different conditions (both containing DNA and apo form) were obtained from these experiments. The observed space groups were *P* 6_4, *C* 121, *P* 4_32_12 and *P* $2_12_12_1$. Importantly, many times crystals were grown exclusively in the DNA-containing solution. However, structure solution invariably showed that the crystals contained the apo form of the protein. It is possible that the length of the DNA fragments is not compatible with crystallization, and that minor variations in the extremes of the DNA sequences are necessary.

We present a Reg_{576} structure obtained with remarkable differences to the structure described above (PDB code: 6GYG). In the novel Reg_{576} structure, the asymmetric unit (ASU) contains two dimers and the crystals belongs to the space group *P* 3_2 1 2. The contacts between the two dimers are formed between the H2 of one dimer with the H3 of the other dimer. The overall structure is shown in **FIGURE 33A**.

We next wondered whether this relative position of the two dimers in the ASU may be representative for the binding to DNA. The length of regRE sequence is 7bp, with a 9bp separation between the start of the two repeats. In B-DNA, which is the most common double helical structure found in nature, the double helix is right-handed with about 10-10.5bp per turn. Reg_{576} binds DNA through its β-sheets and thus, they need to be positioned at an angle of about 45° in respect to the direction of the ds-DNA, which will also be the direction in which the two dimers are aligned. We compared the tetramer observed in the *P* 3_1 1 2 structure with a tetramer model of Reg_{576} based on the DNA-bound Arc structure was constructed by Val-Calvo *et al.*[268] (**FIGURE 33B**)

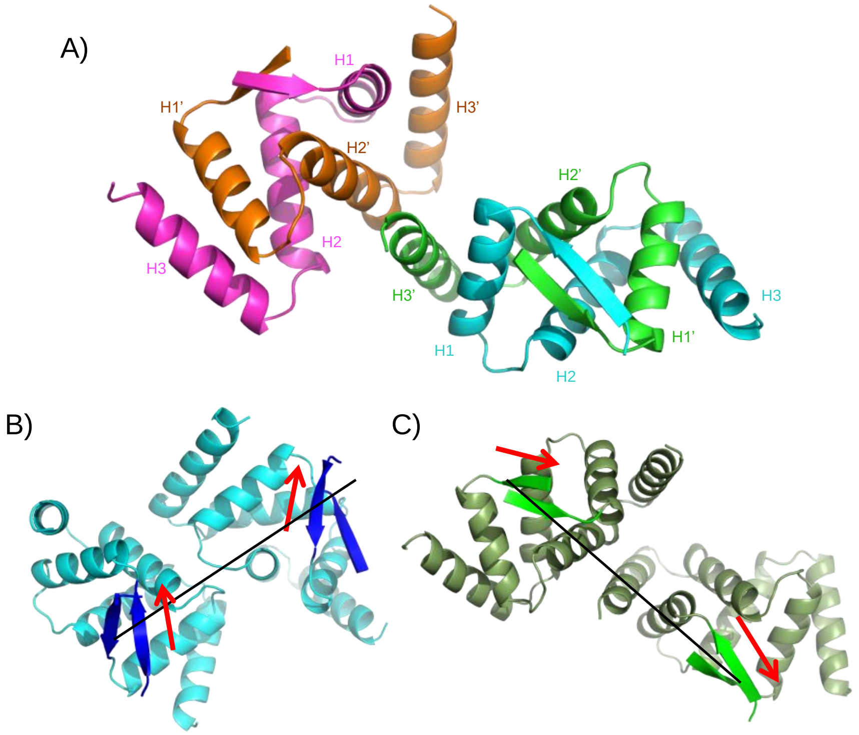


FIGURE 33: **Comparison of the Reg$_{576}$ structures. A)** Cartoon representation of tetramer Reg$_{576}$ structure, colored by chain. Tetrameric model of deposited Reg$_{576}$ (**B**) and tetrameric Reg$_{576}$ obtained in this work (**C**). β-sheets are highlighted by different colors of the overall structure as their position is critical for DNA binding. Red arrows indicate the direction of the β-sheets.

By comparison of the model based on Arc-DNA structure and the structure obtained in this work, we realized the angle between the β-sheets is remarkably higher in the tetrameric Reg$_{576}$ than in the model. This does not necessarily mean that the structure obtained is not representative of how Reg$_{576}$ may bind to DNA. We are comparing it to a model, which we do not have any certainty that is a close representation of how it actually binds. The model comes from a structure of a protein that may have a similar binding based on the sequence length and separation of its DRs. However, they do differ in some other features, like the presence of the extra α-helix, which is also present in Mnt, SSV-RH and NE0241 but none of their structures bound to DNA has been solved. On the other hand, the condition in which the tetrameric form of Reg$_{576}$ was obtained did not contain DNA, so the disposition of the two dimers observed has not been induced by the presence of DNA.

Still, we do not have enough information to decide which of the structures may be closer to how binding occurs between Reg$_{576}$ and DNA. Unraveling the structure of Reg$_{576}$ together with DNA would bring conclusive evidence on this matter. Another strategy should be followed for this purpose. One possibility may be to test different ds-DNAs, those that we have used in this work are at least 20bp in length and the reason for not having obtained any co-crystals with them may be due to fitting troubles in the crystallographic grid. The sequence length of DNA can be decreased up to 16bp.

It may be interesting to force the binding by cross-linking, which consists on chemically joining two or more molecules by a covalent bond. Cross-linkers or cross-linking reagents contain two or more reactive ends that are able to chemically attach to specific functional groups on proteins or other macromolecules. However, the reactivities of most cross-linkers benefit interactions between proteins, which can pose a problem when attempting to cross-link DNA to proteins. To overcome this obstacle, DNA can be synthesized with amines or thiols attached to specific bases, and so, amine or sulfhydryl-reactive cross-linkers would be used on that case. Synthesizing a ds-DNA containing the DR of regRE with these chemical modifications may be a possibility for the future.

C 3.5. CONCLUSIONS

i. Reg$_{576}$ binds to the DR dual heptamer 5´-TTATCCC-3´ as we have seen by testing different length ds-DNAs in SEC. We have named this heptamer regRE.

ii. Reg$_{576}$ does not show affinity for DNAs containing a single regRE, indicating a high degree of cooperativity unlike other RHH proteins.

iii. The middle region of the DR of regRE has also been discarded as the binding region of Reg$_{576}$, reaffirming Reg$_{576}$ binds to the DR of regRE.

iv. Reg$_{576}$ is a RHH protein with an additional α-helix, as is the case for the repressors Mnt, SSV-RH or NE0241.

v. Lysine 46 is located in the loop between the α-helix2 and α-helix3 and it is highly conserved in RHH proteins containing an extra α-helix.

vi. However, binding assays using mutated Reg$_{576}$46K>A also have shown binding to the promoter region, meaning it is not a crucial residue for protein stability or binding to DNA.

vii. We believe that α-helix3 has an important role in the stabilization of Reg$_{576}$ dimer to dimer interface. We think it would it be interesting to further study the effect of this α-helix in the overall Reg$_{576}$ structure and promoter binding.

viii. Co-crystallization assays have been performed to obtain the structure of Reg$_{576}$ in complex with DNA based on the results obtained by SEC. Many different conditions were tested, resulting in many different crystalline forms.

ix. Nevertheless, no Reg$_{576}$ structure in complex with DNA has been obtained.

x. One of the crystals we found interesting belonged to the $P3_212$ space group and contained two dimers in the ASU. The contacts between the two dimers are formed by the H2 of one dimer and the H3 of the other dimer.

xi. We have compared this tetrameric form of Reg$_{576}$ with a previously published model of Reg$_{576}$.[268] This model was generated based on the structure of Arc repressor bound to DNA as they may have a similar DNA binding manner.

xii. We have focused on the β-sheets to do the comparison, as they are required to be positioned at an angle of 45° in respect to the direction of the ds-DNA.

xiii. The model and the tetrameric form of Reg_{576} do not coincide but we do not possess sufficient evidence to conclude which one of them may have more in common with the actual biding manner of Reg_{576} to DNA.

xiv. Our results suggest that a different strategy should be tried in order to obtain the structure of Reg_{576} in complex with DNA.

CHAPTER 4: P34$_{pLS20}$, Cell adhesion of pLS20

C 4.1. OBJECTIVES

Cell adhesion is a very important step in the process of bacterial conjugation. $P34_{pLS20}$ may have a very important role in the adhesion of the donor and acceptor cell, as it may be the responsible for one bacterium to attach another one. It is a large protein that it is predicted to contain a thioester domain, which is a domain involved in covalent adhesion. Thus, this protein may represent a viable target for the disruption of the attachment between cells, thereby inhibiting conjugation. In order to be able to use $P34_{pLS20}$ as a valid target for inhibition of ARGs propagation, knowledge about its structure is highly valuable.

These are the objectives we set for this chapter:

- Confirming $P34_{pLS20}$ is a TIE protein containing a TED by structure elucidation of the domain.
- Unraveling the importance of the thioester bond formation by mutating one of the residues involved and checking its structural implications.
- Analyzing the functional effect of the mentioned mutation in the conjugative process.

C 4.2. INTRODUCTION

One of the most critical steps in colonization and maintenance of the infection is the adhesion of microbes to their target molecules. Bacterial adhesins bind either directly to integral host cell surface components, like integrins or carbohydrates, or they interact with components of the extracellular matrix that result in indirect binding to receptors on the host cell surface.[284] Bacteria express and display a variety of surface proteins, which are subject to environmental stress and therefore must possess significant inherent stability. The interactions described so far in literature are mostly non-covalent and frequently require extensive intermolecular interfaces and multivalent binding. The discovery of internal thioester bonds in the pilus tip adhesin Cpa from Gram + human pathogen *Streptococcus pyogenes* was the first time it was shown that some kind of covalent adhesion could also play a role in adhesion processes.[285,286]

Internal thioester bonds generate covalent links between amino acids side chains. They are found in subunits of pili and other adhesins in Gram + bacteria.[287] Intramolecular isopeptide bonds[288] and ester bonds also belong to the same group of proteins.[289] Whereas intramolecular isopeptide or ester bonds improve protein stabilization, thioester bonds establish covalent adhesion between bacterial surfaces and nucleophilic groups on host tissue targets. Confirmation of the accessibility and reactivity of the thioester was provided by the observation of a covalent bond formation between a Cpa thioester domain and the small molecule nucleophile spermidine.[285] More importantly, binding of *S. pyogenes* to mammalian cells *in vitro* was inhibited with a Cpa variant lacking the thioester domain, providing evidence for the importance of the binding in bacterial adhesion.[286]

Thioester bonds were observed in complement proteins C3 and C4 already in 1997 and were not associated with bacterial adhesion until 2005. These bonds are formed between Cys thiol group and Gln amide group and occur in thioester domains (TED). They probably form due to a favorable environment during protein folding. TEDs are nearly always located at the N-terminal of the protein and contain secretion signals and a LPTXG cell-anchor motif in the C-terminal part of the protein. Most TED-containing proteins also contain isopeptide and/or ester domains, arranged as tandem repeats. These proteins are known as TIE (thioester, isopeptide and ester) proteins. Fibronecting-binding repeats and proline-rich regions are some other domains usually related with TEDs. TIE proteins are increasingly prevalent in surface-associated proteins identified in Gram + bacteria, including some clinically relevant pathogens, such as *Streptococcus pneumonia*, the most frequent cause of pneumonia in adults.

Discovery of TEDs in Cpa prompted further studies searching for similar domains in other proteins. By PSI-BLAST searches, hundreds of potential hits were obtained, which suggested that TIE proteins might be highly abundant. However, they may be very diverse as far as their domain composition and sequences are concerned. By MSAs of up to now characterized TEDs, it is believed they could be divided into two structural classes.[290] Nonetheless, based on the quick increment in genetic information, other structurally distinct TED classes may emerge.

Some Class I TEDs were structurally characterized in 2015, such as a *Clostridium perfringens* TED (PDB code: 5A0G) or SfbI from *Streptococcus pyogenes* (PDB code: 5A0L)[290]. However, there was not much information about the structures of Class II TEDs until recently when three different structures were published by Miller *et al.*[291] Thanks to these newly characterized structures, a wealth of information on the differences between Class I and Class II TEDs is now available. Firstly, Class I TEDs possess an N-terminal indel of 15-20 amino acids that is not present in Class II. Secondly, Class II appears to have an extended C-terminal indel absent in Class I, which results in a 30% mass increase. However, in both Class I and II, the thioester bond forming Cys and Gln residues are identified in a [YFL]CΦζ amino acid motif and a weak ΦQζΦΦ motif (where Φ is any hydrophobic and ζ is any hydrophilic residue) respectively. A third motif, TQXXΦWXΦXζ, was also observed in all Class I TEDs. Similarly to Class I, Class II TEDs have conserved Gln and Trp residues that are positioned on α-helix 2, located directly adjacent to the thioester bond.[285] The Gln/Trp motif in Class II TEDs differs substantially in length from the TQXXΦWXΦXζ motif described for Class I and was redefined to be TQXXΦW, evidencing the lack of conservation after the Trp.[291] Topology diagrams for Class I TED and Class II TED are presented in **FIGURE 34**, schematically representing all differences and similarities between both classes. Furthermore, structural bases are shown in **FIGURE 35**. In both figures, unique features of each class are highlighted by different colors.

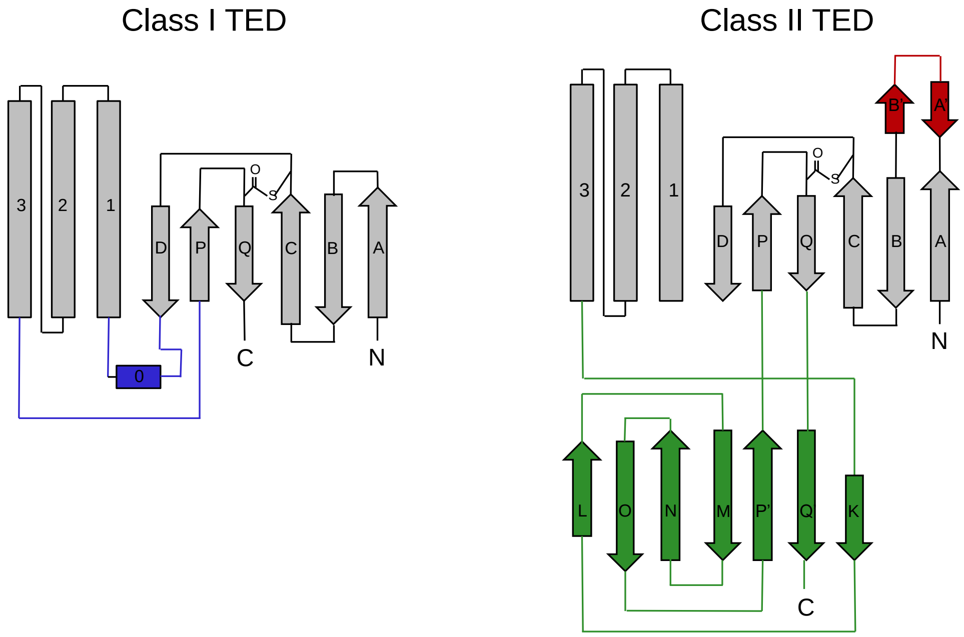


FIGURE 34: **Topology diagrams of SfbI-TED and SaTIE-TED, representatives of TED Classes I and II, respectively.** Secondary structures are labeled with equivalent numbers (α-helices) and letters (β-strands) to reflect features that are conserved in both classes. Blue represents the unique features of TED Class I and red and green indicate the unique features of TED Class II. Thioester bonds are also represented.

The upper lobes of the Class II TEDs correspond to the canonical TED fold, which is composed of six-stranded antiparallel β-barrel and a three-helix bundle. The most significant difference between Class I and Class II TEDs is the replacement of a linker of approximately 10 residues between α-helix 3 and β-strand P in Class I TED with a seven-stranded β-sandwich, composed of 75 residues. Another difference is that in Class I, a α-helix (α0) links β-strand D and α-helix 1, which is not present on Class II. α0 coincides with the N-terminal indel identified at the N terminus by the sequence alignments of TED domains. In both Class I and II, the thioester bond-forming Cys is located in β-strand C, whereas the Gln counterpart is contributed by β-strand Q. The β-sandwich formed between β-strands Q and P is formed by a tangled β-hairpin that complements the β-barrel in the N-terminal, and thus, forms a slipknot-like structure.

There are two structural elements that restrict the receptor access to the thioester bond. On the one hand, the loop between β-strands A and B, which is located between the β-barrel and helical subdomains (represented in red in **FIGURE 35**). As mentioned before, this loop is approximately 15 residues longer in Class II TEDs than Class I TEDs, hindering the access to the thioester bond. The other restriction concerns the depth of the cleft, which is over 8Å deep. To understand how this feature obstructs the accessibility to the thioester bond, the depth of this pore is longer than the depth of an Arg side-chain, which has one of the longest side chains of all amino acids. Hence, it seems that the interaction between the thioester bond and its cognate receptor needs a previous conformational change of the surrounding structure. So as to undergo this change, some recognition events around the specificity loop may be required. Recent evidence by atomic force microscopy (AFM) of the C-terminal TED from Cpa supports the idea that a structural rearrangement is required for bond formation between a TED and its cognate receptor. [292]

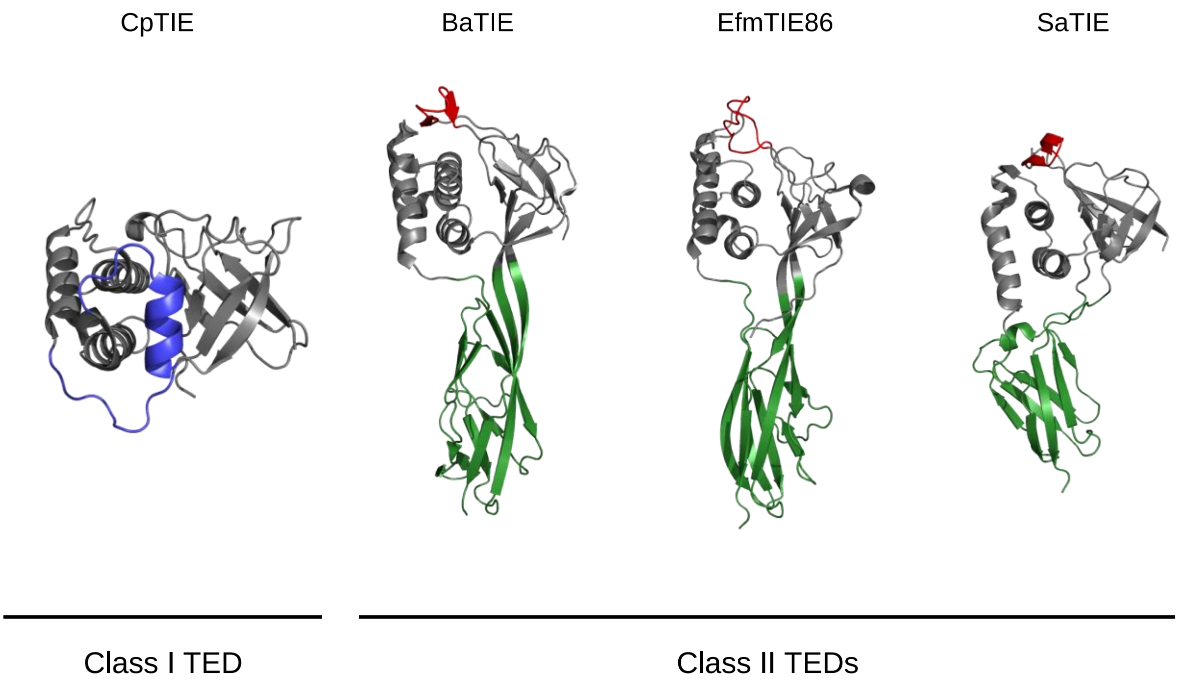


FIGURE 35: **Structural basis of TED classification.** One Class I TED (CpTIE, PDB code: 5A0G) versus three Class II TEDs (BaTIE, PDB code: 6FWV; SaTIE, PDB code: 6FX6; EfmTIE86, PDB code: 6FWY) shown in cartoon representation. Colors represent the structural elements that determine both classes. Blue shows the α0 helix that links β-strand D and α-helix 1 that is not present in Class II TEDs. Red and green represent the indels that are exclusive from Class II TEDs; The N-terminal indel and the β-sandwich forming a slipknot structure, respectively. Colors that represent the structural elements coincide with those in the topology diagram.

There is not enough evidence to determine whether the structural differences in TED classes have a functional significance in their reactivity, interaction with receptors or other biological roles. It is worth mentioning that some genera of Gram + bacteria encode either Class I or Class II TEDs but not both. For example, *Streptococcus* thioester proteins restrictively belong to Class I TEDs, while *Enterococcus* and *Bacillus* encode exclusively Class II.

Covalent binding of bacteria to its targets had only been described in mammalian complement family (C3, C4 and some other related proteins),[293,294] which function by covalently tagging pathogens to phagocytosis. Thus, covalent binding in bacterial adhesion is a unusual event of protein chemistry convergent evolution.[295,293] Nonetheless, it is remarkable that bacterial TEDs and complement proteins are unrelated both sequentially and structurally. Complement proteins are multi-domain constructs that need a proteolytic cleavage for their activation. When cleaved, they undergo a conformational change by which the thioester gets exposed.[296] Subsequently, the thioester reacts nonspecifically with nucleophiles like water and nucleophilic moieties on bacterial surfaces.[297] Contrarily, bacterial TEDs that have been characterized comprise only a single domain, and do not depend upon proteolytic activation. Moreover, they show high selectivity, as demonstrated by SfbI-TED binding to fibrinogen.[290] It remains unknown how access to and reactivity of bacterial thioesters are regulated. Therefore, it seems that a parallel, convergent evolution has led to the occurrence of TED domains that function as a molecule for irreversible binding both in bacteria and complement proteins.

In this chapter, we have studied the protein encoded by gene *p34* of pLS20 plasmid and found out that it is a TED-containing protein belonging to Class II. It is a very large protein, of about 800 residues and a M_W of 84.4kDa. It is predicted to have a transmembrane domain. Also, it may have CnaA and CnaB domains, which are Ig-type domains that are used for the dissipation of mechanical energy by unfolding and refolding of isopeptdie bond-delimited polypeptide loops in Gram + pili. Despite its dimensions, typically 30-40Å thick, the assembly is maintained by covalent linkages. A model of the protein proposed by David Abia and Wilfried J.J. Meijer is shown in **FIGURE 36**.

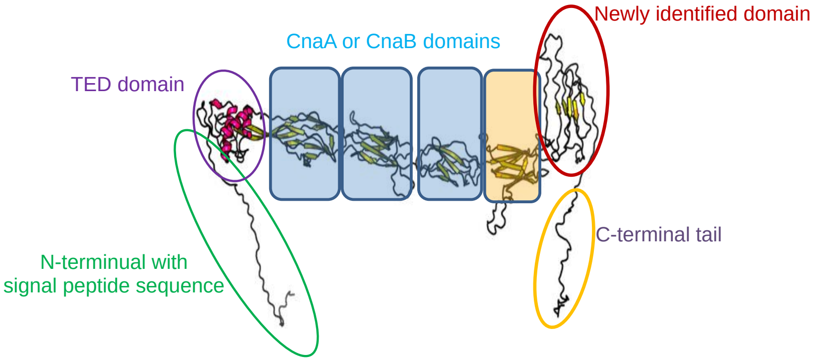


FIGURE 36: **Representation of a $P34_{pLS20}$ model.** Predicted different domains are highlighted using different colors. Protein is expected to have a length of 220-230Å.

Disruption of inhibition to suppress ARG propagation has been tested using different targets such as relaxases or complementary proteins. Nevertheless, proteins containing TED domains have never been targeted and provide a world of new possibilities. Gathering structural information about $P34_{pLS20}$TED would be extremely helpful for inhibiting cell-to-cell connections that take place in the conjugative process for the purpose of ARG transmission.

In this work, we have structurally characterized the TED of $P34_{pLS20}$, confirming it is a TIE protein containing a Class II TED. Also, we have determined the structure of a single point mutant of $P34_{pLS20}$, changing one of the residues involved in the thioester bond, and checked how the bond formation results impaired. By functional assays, we have proven not only that $P34_{pLS20}$ is required for a successful conjugative event, but also that thioester bond is necessary for this to happen.

C 4.3. MATERIALS AND METHODS

C 4.3.1. Production of $P34_{pLS20}$[35-285], $P34_{pLS20}$[35-285]C68S and $P34_{pLS20}$FL

Cloning of *$p34_{pLS20}$* and *$p34_{pLS20}$C68S and p34FL* was done in Wilfried JJ Meijer's lab in the CBMSO (CSIC, UAM). Besides FL form, two different *$p34_{pLS20}$* constructs were designed, all of them using the pET28b expression vector: first clone including residues 35-285 ($P34_{pLS20}$[35-285]) and another one including residues 35-269 ($P34_{pLS20}$[35-269]). In case of the mutant protein, named *$p34_{pLS20}$C68S* , it included residues 35-285 but had a serine in position 68 instead of a cysteine. The coding region was placed in frame upstream a region coding for six histidine residues in all cases.

Expression of *$p34_{pLS20}$FL* was done as explained in §2.1"Protein cloning and expression protocol" but expression of the other two constructs was slightly different. They were induced by 1mM IPTG incubated at 20°C for 20h instead of at 37°C, performed in Wilfried JJ Meijer's lab. $P34_{pLS20}$[35-269] could not be obtained when grown under these conditions. However, expression of the resulting $P34_{pLS20}$[35-285] and $P34_{pLS20}$FL resulted in accumulation of a protein with the expected size. In addition, a Selenomethionine (SeMet) derivative of $P34_{pLS20}$[35-285] was also produced by using "SelenoMethionine Medium Complete" kit (Molecular Dimensions).

Purification was performed for $P34_{pLS20}$[35-285], $P34_{pLS20}$FL and the SeMet derivative $P34_{pLS20}$[35-285] as described as §2.2"Chromatography techniques and Protein purification protocol". A second purification based on particle size, which is 26kDa in case of $P34_{pLS20}$[35-285] and 85kDa in case of the $P34_{pLS20}$FL, was done using ProteoSEC 3-70HR and ProteoSEC 6-600HR columns, respectively, with 500mM, 20mM TRIS pH8 buffer. Enough quantity of pure protein was obtained for both $P34_{pLS20}$[35-285] and SeMet $P34_{pLS20}$[35-285], but not for the FL form of the protein.

C 4.3.2. Crystallization of $P34_{pLS20}$[35-285] and $P34_{pLS20}$[35-285]C68S and Crystallization of $P34_{pLS20}$[35-285] with fragment screenings

Pure $P34_{pLS20}$[35-285] was concentrated to 10mg ml^{-1} in a buffer containing 500mM NaCl, 20mM TRIS pH 8.0 over an Amicon® Ultra 15mL centrifugal filter with a cutoff of 10kDa. Crystallization experiments were setup using the vapor-diffusion method on both sitting and hanging drop at 18°C. Initial crystals were obtained after initial trials with commercial screenings that are described in §2.4"Crystallization techniques". Promising conditions were scaled up. For the final crystals, 1µl of protein + 1µl of crystallization buffer (0.2M Sodium acetate, 30% PEG 8000, 0.1M Sodium cacodylate pH 6.5) were used against 500µl of crystallization buffer in the reservoir. Selenomethionine crystals were obtained in the same conditions.

Pure P34$_{\mathrm{pLS20}}$[35-285]C68S was also concentrated to 10mg ml^{-1} being its buffer 500mM NaCl, 20mM TRIS pH 8.0. Crystals were obtained with the same method utilized for the native protein. Yet, in this case the crystallization buffer was 2M Ammonium sulfate, 0.1M Hepes pH 7.5. Native and mutant crystals were fished and cryo-cooled in liquid nitrogen for X-ray collection.

Also, P34$_{\mathrm{pLS20}}$[35-285] was co-crystallized with several solvents using a fragment screening library (Maybridge Ro3 2500 Diversity Fragment Library). 100nl of protein at 10mg ml^{-1} in a buffer containing 500mM NaCl, 20mM TRIS pH 8.0 + 100nl crystallization buffer (0.2M Sodium acetate, 30% PEG 8000, 0.1M Sodium cacodylate pH 6.5) + 20nl of fragment from a 200mM stock (with a final concentration of 20mM) were equilibrated against 50µl crystallization buffer. Crystals started to appear after 3 days of incubation time and were harvested and cryo-cooled in liquid nitrogen for X-ray collection.

C 4.3.3. Diffraction, data Processing of SeMet-P34$_{\mathrm{pLS20}}$[35-285] and P34$_{\mathrm{pLS20}}$[35-285]C68S and Structure comparison

Data collection was performed at ALBA Synchrotron Light Source on the BL13-Xaloc beamline[196]. The structure of P34$_{\mathrm{pLS20}}$[35-285] was determined using Se-SAD phasing, while P34$_{\mathrm{pLS20}}$[35-285]C68S was determined by MR using P34$_{\mathrm{pLS20}}$[35-285] as a model. Collected data were processed with XDS program.[298] Intensities were scaled and truncated with Aimless.[299] In case of P34$_{\mathrm{pLS20}}$[35-285], the selenium positions were determined by using HKL2map. Model building was performed by ARP/wARP.[300] The model was refined automatically with Phenix.refine[281] and REFMAC5 from CCP4 suite[201] and also manually with COOT.[282] In case of P34$_{\mathrm{pLS20}}$[35-285], PHASER[168] from the CCP4 suite was used for the MR.

Figures were prepared using PyMOL (The PyMOL Molecular Graphics System version 2.3 Schrödinger, LLC).[301] Superpositions of P34$_{\mathrm{pLS20}}$[35-285] were performed using the align function in PyMOL. F_o-F_c maps were prepared with CCP4i.[201] RMSDs for structure comparisons were calculated by SSH option in Coot.[282]

C 4.4. RESULTS

C 4.4.1. Purification trials of P34$_{pLS20}$FL and Peptide Mass Fingerprinting assay

We did not succeed in obtaining sufficient amounts of pure P34$_{pLS20}$FL. Instead, we observed a contamination of 55kDa in addition to the 84kDa-sized gel band corresponding to P34$_{pLS20}$FL (left electrophoresis gel in **FIGURE 37A**). In order to separate these two proteins, we decided to perform a SEC using a ProteoSEC 6-600HR (Generon) column. However, traces of the 55kDa protein persisted (right electrophoresis gel in **FIGURE 37B**).

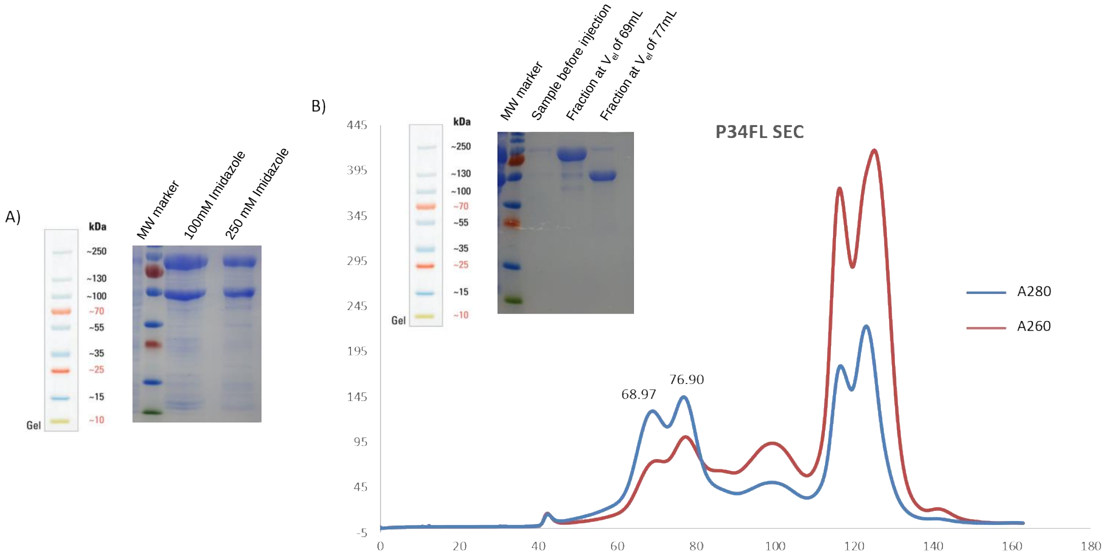


FIGURE 37: **P34$_{pLS20}$FL purification and SEC electrophoresis gels. A)** 12% electrophoresis gels from His-trap purification, showing fractions obtained in 100mM and 250mM imidazole, from left to right. **B)** SEC elution profile from a ProteoSEC 6-600HR column, plotting the absorbances at 280nm and 260nm (in blue and purple, respectively) against the V$_{el}$. The inset shows a 12% electrophoresis gel of protein sample before injection and maximum fractions from peak at V$_{el}$ of 69 and 79ml, from left to right.

To elucidate whether the 55kDa protein consisted on a degradation of the FL protein or a contamination, we decided to perform a protein mass fingerprinting (PMF) assay. For that we sent a sample of the fragment to the Proteomics and genomics service from Centro de Investigaciones Biológicas for analysis by MALDI-TOF/TOF. Results are displayed in TABLE **9**. The best results (WP_01603195.1, WP_033881018.1, GAK78262.1, WP_088272629.1, WP_072176328.1) sequences were checked and we realized they all belonged to TED-containing proteins and aligned perfectly to P34$_{pLS20}$FL, revealing the 55kDa fragment corresponded to a degradation product.

Importantly, the degradation product aligned with the C-terminal part of the protein, meaning degradation occurs at the N-terminal part. This result is in line with the fact that 55kDa-sized fragment bound the nickel column, as the histidine tag is in the C-terminal part of the protein. Based on the M_w and the model proposed by David Abia and Wilfried J.J. Meijer (**FIGURE 36**), the degradation product may lack the N-terminal signal peptide and the TED.

TABLE 9: **Mass fingerprinting results for the 55kDa protein obtained in P34$_{pLS20}$FL purification.** Most relevant hits by PMF assay are displayed, ordered by relevance and showing protein accession numbers, descriptions and M_{WS}.

#	Protein accession number	Mass (Da)	Score	Description
1	WP_013603195.1	84270	370	hypothetical protein [Bacillus subtilis]
2	WP_033881018.1	86892	367	hypothetical protein [Bacillus subtilis]
3	GAK78262.1	87005	367	hypothetical protein BSMD_001580 [Bacillus subtilis Miyagi-4]
4	WP_088272629.1	85625	366	hypothetical protein [Bacillus subtilis]
5	WP_072176328.1	85654	208	hypothetical protein [Bacillus subtilis]
6	WP_038429014.1	18557	45	hypothetical protein [Bacillus subtilis]
7	WP_041850798.1	18340	42	MULTISPECIES: hypothetical protein [Bacillus]
8	WP_076458762.1	18410	42	hypothetical protein [Bacillus subtilis]
9	WP_015417613.1	11897	41	MULTISPECIES: tRNA-binding protein [Bacillus]
10	WP_019713060.1	31438	41	energy-coupling factor ABC transporter ATP-binding protein [Bacillus subtilis]
11	WP_014419272.1	46217	40	MULTISPECIES: pyrimidine-nucleoside phosphorylase [Bacillus]
12	WP_015715262.1	16411	39	MULTISPECIES: NUDIX hydrolase [Bacillus]
13	WP_029726552.1	16443	39	NUDIX hydrolase [Bacillus subtilis]
14	WP_049141405.1	16441	39	NUDIX hydrolase [Bacillus subtilis]
15	WP_063336084.1	16499	39	NUDIX hydrolase [Bacillus subtilis]
16	WP_100506707.1	16516	39	NUDIX hydrolase [Bacillus subtilis]
17	CUB59692.1	12313	39	Xylulose kinase [Bacillus subtilis]
18	WP_106074078.1	16455	39	NUDIX hydrolase [Bacillus subtilis]
19	WP_007409920.1	28210	38	MULTISPECIES: tRNA1(Val) (adenine(37)-N6)-methyltransferase [Bacillus]
20	WP_004264717.1	28196	38	MULTISPECIES: tRNA1(Val) (adenine(37)-N6)-methyltransferase [Bacillus]

It seems that P34$_{pLS20}$ is not a stable protein and that the production and structure determination of the FL protein may therefore be challenging. An alternative approach would be to solve the structure of the different domains and reconstitute a model of the full protein from these partial structure.

C 4.4.2. Crystallization of P34$_{pLS20}$[35-285] and P34$_{pLS20}$[35-285]C68S

P34$_{pLS20}$[35-285] crystals belong to P 4_12_12 space group with unit cells of a=48.799, b=48.799 and c=204.48 and α=90, β=90 and γ=90, at a resolution of 1.57Å The calculated Matthew’s coefficient was of 2.71Å3/Da, which is consistent with a single molecule in the ASU.

P34$_{pLS20}$[35-285]C68S crystals belong to *P* $4_3 2_1 2$ space group with unit cell dimensions of a=74.90, b=74.90, c =243.41 and α=90, β=90, γ=90, at a resolution of 2.49Å. In the case of the mutant, 2 molecules were calculated to fit in the ASU by a Matthew's coefficient of 2.53Å3/Da.

The statistics of the processing of the X-ray diffraction data, the refinement of the final structure and its validation of P34$_{pLS20}$ and P34$_{pLS20}$C68S are summarized in TABLE **10**.

TABLE 10: **X-ray processing and refinement statistics for P34$_{pLS20}$ and P34$_{pLS20}$C68S.**

Data processing statistics	**P34$_{pLS20}$**	**P34$_{pLS20}$C68S**
Space group	*P* $4_1\ 2_1\ 2$	*P* $4_3\ 2_1\ 2$
Unit-cell parameters (Å)	a=48.799 b=48.799 c=204.48	a=74.90 b=74.90 c=243.41
Unit-cell angles (°)	α=90 β=90 γ=90	α=90 β=90 γ=90
Resolution range (Å)[a]	1.50-51.12	2.49-48.61
No. of unique reflections	41022	24757
Completeness (%)	99.99	95.8
Redundancy	24.4	3.1
Mean I/σ(I)	11.8	8.9
R_{meas} (%)[b]	0.155	0.107
Refinement statistics		
R_{work}[c] (%)	18.12	24.78
R_{free}[d] (%)	20.26	31.10
Ramachandran		
Favored (%)	99.19	92.02
Disallowed (%)	0.81	1.60
R.M.S.D.		
Bond lengths (Å)	0.005	0.0095
Bond angles (°)	0.778	1.16
Chirality		
Mean B value (Å^2)	25.6	45.4

[a]Numbers in parentheses represent values in the highest resolution shell.

[b]Rmeas=$\sum$hkl $[N/N-1]^{1/2} \sum_i$ |Ii(hkl) - <I(hkl)>| / $\sum_{hkl} \sum_i$ Ii(hkl) where N is the multiplicity of a given reflection, Ii(hkl) is the integrated intensity of a given reflection, and <I(hkl)> is the mean intensity of multiple corresponding symmetry-related reflections.

[c]Rwork=$\sum$ ||Fobs| - |Fcalc|| /$\sum$ |Fobs|, where |Fobs| and |Fcalc| are the observed and calculated structure factor amplitudes, respectively.

[d]R_{free} is the same as R_{work} but calculated with a 5% subset of all reflections that was never used in refinement.

C 4.4.3. Structures of P34$_{pLS20}$TED and P34$_{pLS20}$TEDC68S

The structure of the construct of P34$_{pLS20}$ containing residues 35-285 was found to be structurally similar to the TED-containing group of proteins, as predicted by the sequence homology. We will refer to this domain as P34$_{pLS20}$TED from now on. More specifically, it belongs to Class II types of TED domains as judged by the characteristics they share. The most evident Class II characteristic is the 7-stranded β-sandwich forming a slipknot structure, which replaces an approximately 10 residues linker between α-helix 3 and β-strand P in Class I TEDs. This domain is composed of about 80 residues that is missing in Class I TEDs. In addition, P34 $_{pLS20}$TED contains an indel between β-strands A and B. Interestingly, this indel is composed of two β-strands (β-strand A' and B'). Finally, it lacks the α0 helix that links β-strand D and α-helix 1 and an N-terminal indel of approximately 15-20 residues that are present on Class I TEDs.

As represented in **FIGURE 38**, the upper lobe of the P34$_{pLS20}$TED corresponds to canonical Class I TED structure, encompassing a five-stranded β-barrel and a three-helix bundleresidues. A thioester bond is formed between Cys68 from β-strand C and Gln254 from β-strand Q. Both of them are found in the conserved motifs [YFL]CΦζ and ΦQζΦΦ, respectively: Cys68 is followed by an Ile and an Asp, while Gln254 is located in a Tyr, Gln254, Arg, Leu and Met. Remarkably, the Gln/Trp motif cannot be found in P34$_{pLS20}$TED. By superimposition with SaTIE, BaTIE and EfmTIE86, this motif would correspond to a VALNNW, in which only the W (Trp125) is conserved. Nevertheless, mutagenesis of Cpa-TED demonstrated that neither the Gln nor Trp of this motif are essential for thioester bond formation.[285] Our results are in line with this finding, as we can see in **FIGURE 38C**, there is no impairment of the thioester bond even if only the W is maintained in the TQXXΦW motif present in other Class II TEDs. These results together suggest that this motif may not be preserved in other TEDs that have not yet been structurally characterized. β-strands Q and P form an extended and twisted β-hairpin that loops back through the N-terminal lobe to complement the β-barrel subdomain, forming a slipknot-like structure, as it is the case in the other Class II TED structures.

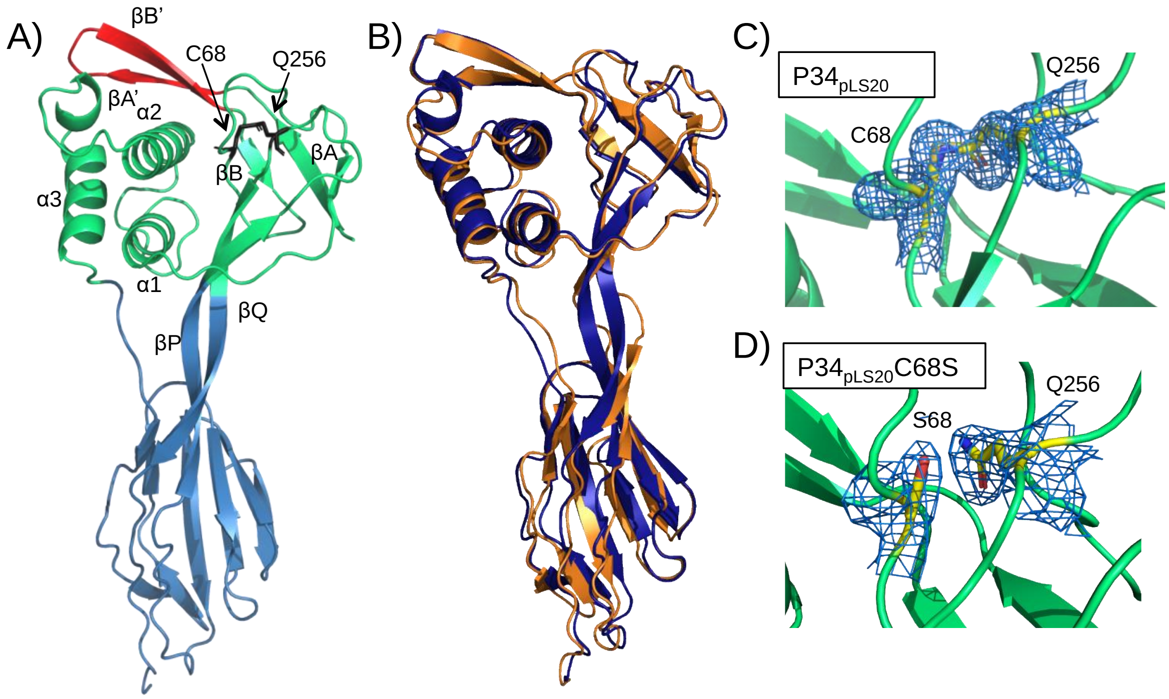


FIGURE 38: **P34TED$_{\text{pLS20}}$ structural details. A)** Cartoon representation highlighting the class-defining indels: the extended C-terminal indel (shown in red) and the slipknot structure (shown in blue). Residues involved in thioester bond are represented as black sticks. **B)** Superposition of P34TED$_{\text{pLS20}}$ (blue) and P34$_{\text{pLS20}}$C68S (orange). Zoom in the thioester bonds of P34TED$_{\text{pLS20}}$ (**C**) and P34TED$_{\text{pLS20}}$C68S (**D**). Residues involved are shown as sticks with 2(F_{obs}–F_{calc}) omit density superimposed.

Accessibility to thioester bond is restricted by β-strands A' and B' as represented in red in **FIGURE 38A**. This region has been hypothesized to be in charge of specificity for binding partners, although there is no supporting experimental evidence.[290]

In case of P34$_{\text{pLS20}}$TEDC68S, we have not detected any substantial differences in its overall structure regarding P34$_{\text{pLS20}}$ wt. However, as shown in **FIGURE 38D**, the thioester bond formation results impaired.

C 4.4.4. Comparison of P34$_{\text{pLS20}}$ with TED proteins

To confirm its high resemblance with Class II TED proteins, we conducted a search by the DALI server. Furthermore, another search was conducted introducing the slipknot-like structure alone. Results are shown in TABLE **11** and TABLE **12**.

TABLE 11: **Structure alignment results for P34$_{\text{pLS20}}$TED.** Different scores used for structure recognition are shown together with the organism to which the hit belongs. Results are sorted by Z and only the twenty most relevant ones are presented.

##	Chain	Z	RMSD	lali	nres	%id	Protein accession number	Organism
1	6fwy-A	13.8	4.1	209	280	18	A0A1A7T0E1	*E. faecium*
2	6fwv-A	13.8	4.5	182	522	15	Q81XH9	*B. antracis*
3	5dcq-E	12.9	2.3	129	200	15	Q93ED6	*Streptococcus equi subsp. equi*
4	4c0z-F	11.7	2.6	125	209	16	S5FVI9	*S.pyogenes*
5	2xid-B	11.5	12.8	137	430	16	Q8GRA2	*S.pyogenes*
6	6fx6-A	11.0	4.3	151	237	15	A0A3F2YM24	*S. aureus*
7	5a0n-A	8.5	3.2	122	214	14	A5MC6	*S. pneumoniae*
8	6fvi-A	6.7	6.4	99	142	10	Q8TEP8	*H. sapiens*
9	2qsv-A	6.6	2.8	87	220	9	Q7MX90	*P. gingivalis*
10	3zmr-A	6.4	2.3	80	469	10	A7LXT7	*B. ovatus*
11	1m1s-A	6.4	2.8	84	109	2	Q23246	*C. elegans*
12	4aq1-A	6.3	7.5	116	721	17	Q45664	*G. stearothermophilus*
13	2wln-A	6.2	3.2	97	228	11	D9N164	*C. perfringes*
14	4uj6-A	6.1	3.6	107	711	16	O68840	*G. stearothermophilus*
15	5ftx-A	5.8	3.3	101	651	16	O68840	*G. stearothermophilus*
16	6fm5-A	5.6	2.4	85	157	8	Q6XBY7	*A. baumannii*
17	4zmh-A	5.6	12.5	100	935	9	Q21JW4	*S. degradans*
18	6n08-A	5.6	9.4	83	488	8	H9XIV6	*Influenza A virus*
19	4ncd-A	5.6	4.1	86	211	6	-	*E. coli*
20	2pn5-A	5.6	12.7	117	1283	9	Q9GYW4	*A. gambiae*

TABLE 12: **Structure alignment results for P34$_{\text{pLS20}}$TED slipknot-like structure.** Different scores used for structure recognition are shown together with the organism to which the hit belongs. Results are sorted by Z and only the twenty most relevant ones are presented.

##	Chain	Z	RMSD	lali	nres	%id	Protein accession number	Organism
1	6fwv-A	10	2.2	98	522	13	Q81XH9	*B. antracis*
2	6fwy-A	9.2	2.4	96	280	16	A0A1A7T0E1	*E. faecium*
3	2qsv-A	6.7	2.6	84	220	10	Q7MX90	*P. gingivalis*
4	6fvi-A	6.6	2.7	87	142	11	Q8TEP8	*H. sapiens*
5	2wln-A	6.5	2.8	92	228	11	D9N164	*C. perfringes*
6	1m1s-A	6.4	2.6	84	109	2	Q23246	*C. elegans*
7	4uj6-A	6.0	2.5	76	711	16	O68840	*G. stearothermophilus*
8	5dfk-A	5.9	2.5	80	185	5	P77188	*E. coli*
9	6kjk-A	5.7	2.9	88	134	6	B2RGZ7	*P. gingivalis*
10	5ftx-A	5.7	2.6	77	651	16	068840	*G. stearothermophilus*
11	3zmr-A	5.6	2.4	77	469	13	A7LXT7	*B. ovatus*
12	5n40-A	5.6	2.7	83	352	13	A0A0U1R0I0	*M. smegmatis*
13	2l3b-A	5.5	3.0	88	130	13	Q8ABM6	*B. thetaiotaomicron*
14	4ncd-A	5.5	3.0	82	211	7	-	*E. coli*
15	6hdp-A	5.4	2.5	77	955	8	Q92630	*H. sapiens*
16	6fm5-A	5.3	2.4	83	157	8	Q6XB47	*A. baumannii*
17	1z91-A	5.3	2.8	84	126	10	034777	*L. interrogans*
18	2mqg-A	5.2	2.6	75	102	8	B5TXC6	*Influenza A virus*
19	1t0p-B	5.2	2.7	74	86	5	P20701	*H. sapiens*
20	3kpt-B	5.2	3.2	92	355	8	Q81D71	*B. cereus*

As we can observe when entering both the whole P34$_{\text{pLS20}}$TED (TABLE **11**) and the slipknot structure alone (TABLE **12**) the results obtained are similar. However, 8 new hits are observed when running the DALI server with the slipknot structure alone (PDB codes: 5DFK, 6KJK, 5N40, 2L3B, 6HDP, 2MQG, 1T0P, 3KPT).

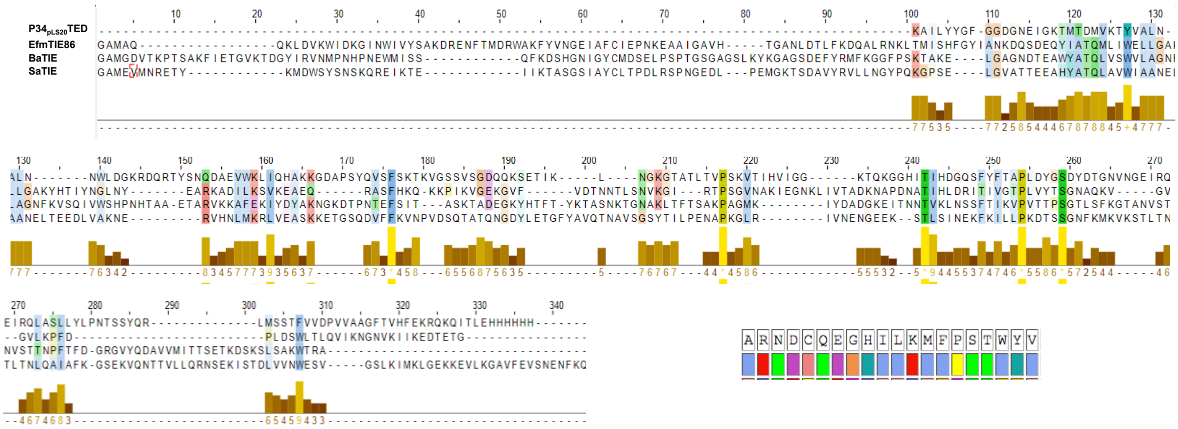


FIGURE 39: **MSA of P34$_{\text{pLS20}}$TED with Class II representatives EfmTIE86, BaTIE and SaTIE.** Conservation scores are shown below the sequences and color legend for each residue is also displayed.

As expected, all three Class II representatives are among the most relevant results (PDB codes: 6FWY, 6FX6, 6FWV). They have pairwise identities of 18-23%, however they exhibit very similar tertiary structures. Searches with DALI protein comparison server mark thioester domain of the *Enterococcus faecium* TIE86 protein (EfmTIE86) and *Bacillus anthracis* TIE protein (BaTIE) to be the closest homologs of P34$_{pLS20}$, having a 17% and 15% of sequence homology, respectively. EfmTIE86 has an RMSD of 3.8785Å and a DALI Z of 15.0, whereas BaTIE has an RMSD of 3.9914Å and a DALI Z of 13.9. The other Class II TED representative, which is the TIE domain of *Staphylococcus aureus* (SaTIE), also shows high similarity with P34$_{pLS20}$, having a RMSD of 3.8245Å and a DALI Z of 11.0. Despite fairly low pairwise identities, the tertiary structures have a high degree of similarity.

Superpositions of P34$_{pLS20}$TED with other Class II TED containing proteins show that the main differences are detected in the specificity loop. The upper lobe seems to be very conserved both in Class I and Class II TEDs. However, the slipknot structure shows higher positional variation and their length is variable. An example of this is SaTIE, whose β-sandwich is shorter than the rest Class II representatives. Furthermore, α-helix 3 is remarkably longer than usual in this structure.

Also, many hits related to Class I TEDs are also obtained, such as PDB codes 4C0Z, 2XID or 5A0N. These hits are restricted to the search with the whole P34$_{pLS20}$TED structure, as the similarity is due to the upper lobe of the structure, which is not included in the slipknot-like structure. Regarding the TED β-sandwich domain, DALI identifies proteins that contain immunoglobulin (Ig)-like folds, which encompasses different types of proteins like chaperons, adhesins or signaling proteins. Superpositions of P34$_{pLS20}$TED and its upper and bottom lobe with some of the hits obtained by DALI are displayed in **FIGURE 40**.

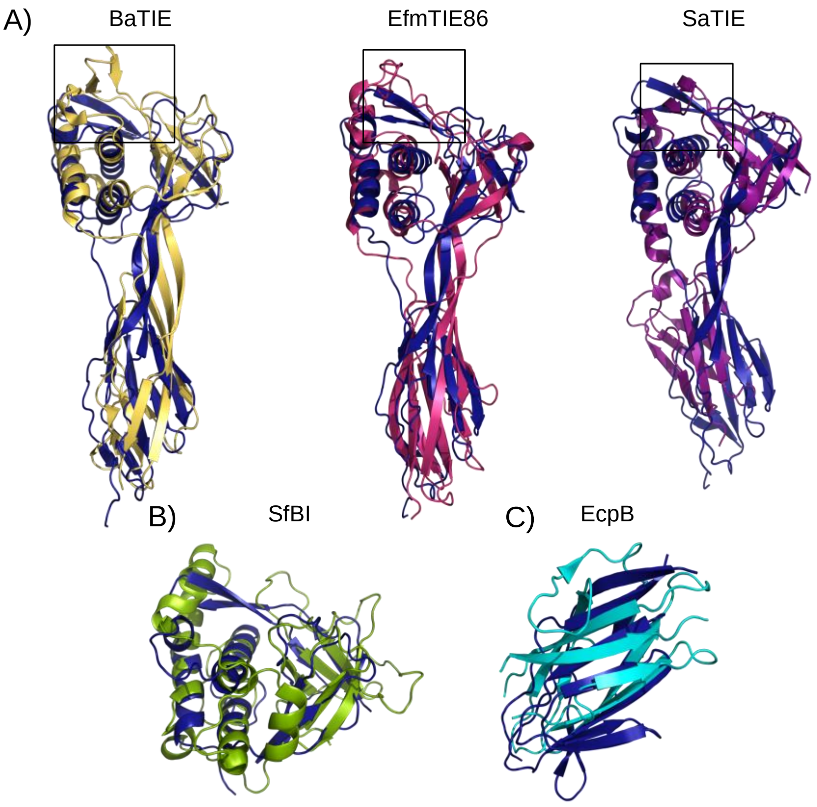


FIGURE 40: **Comparison of full P34$_{pLS20}$TED, P34$_{pLS20}$TED upper lobe and bottom lobe with structurally similar proteins. A)** Superpositions of P34$_{pLS20}$TED structure with up to now characterized Class II TEDs. Specificity loop is highlighted with a square. **B)** Superposition of P34$_{pLS20}$TED upper lobe with SfbI (PDB code: 5A0L). **C)** Superposition of P34$_{pLS20}$TED bottom lobe with EcpB (PDB code: 5DFK). P34$_{pLS20}$TED is represented in dark blue in all cases.

Functional assays were also performed by César Gago at Wilfried J.J Meijer's lab in the Centro Biologia Molecular "Severo Ochoa". Both P34$_{pLS20}$TED and P34$_{pLS20}$TEDC68S were tested. A knock-out of gene *p34$_{pLS20}$* resulted in a 1000x reduction of conjugation efficiency in liquid medium. Furthermore, conjugation efficiency was also diminished with P34$_{pLS20}$C68S mutant, but to a lesser extent.

Our combined functional and structural results indicate that P34$_{pLS20}$ is an important protein in the mating-pair formation. Thioester bond between C68 and Q256 seems to have an essential role in this process, as when this bond is impaired by C68S mutation conjugation efficiency also drops remarkably. As far as we know, P34$_{pLS20}$ is the first TIE protein involved in mating pair formation protein encoded by a conjugative plasmid. Therefore, even though the subdomains of P34$_{pLS20}$TED and other Class II TEDs are similar to already identified and characterized proteins, the combination of its function in mating pair formation, its domain organization and sequential features make P34$_{pLS20}$ unique when compared to homologous proteins with known structures. For example, membrane anchoring is different in P34$_{pLS20}$. Previously described Class II TED representatives contain a LPXTG motif, but the C-terminal domain of P34$_{pLS20}$ does not show evidence for such a motif. It is still unknown how P34$_{pLS20}$ remains attached to the membrane.

Most TED-containing proteins are predicted to contain intramolecular isopeptide and/or ester domains, which are usually present in tandem repeat arrays. Another domain commonly associated with putative TEDs are fibronecting-binding repeats and proline rich regions. Sequence analysis suggests that P34$_{pLS20}$ TED may also be followed by isopeptide or ester domains, as is the case for BaTIE. However, attempts to obtain the structure from the whole P34$_{pLS20}$ were unsuccessful. The protein seems to be unstable as it degrades easily as proven by fingerprinting assays. One of the approaches that could be followed is cloning the different domains of the protein and obtaining the structure of these separately and merging it. A candidate construct for this strategy is the 55kDa degradation fragment obtained when producing the FL protein.

Another interesting continuation of this project would be to identify ligands of P34$_{pLS20}$ that inhibit its function. We have embarked on initial trials using crystallographic fragment screenings. This technique was first used in 1996 to discover small molecule binding sites by soaking organic solvents into crystals to then diffract them by X-rays.[302] By superimposing crystal structures that have been obtained from soaking different ligands, conformational changes, plasticity or chemical complementarity can be inferred. We obtained hundreds of crystals were obtained (some of them are shown in **FIGURE 41**) from which we obtained several dozens of datasets. However, we have not identified ligands in the corresponding density maps. The potentially high number of false positives containing only the protein on its apo form is a known pitfall of this technique. We are now considering ways to deal with the large amounts of data that need to be analyzed.

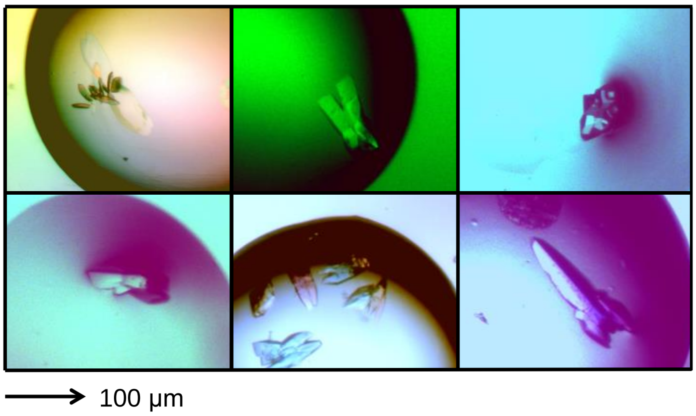


FIGURE 41: **Crystals of P34$_{pLS20}$TED in the presence of different fragments.** Some examples of the crystals obtained with fragments from the Maybridge Ro3 2500 Diversity Fragment Library are shown. All crystals were grown under the same conditions and crystallization buffer. Pictures were taken with an optical microscope using polarized light. Images are shown in scale.

Considering that fragment screening by crystallography did not result in a ligand with affinity to P34$_{pLS20}$, al alternative technique to perform high throughput analysis to detect potential binding partners may be a good approach. For this purpose, plates with immobilized labeled-ligands on the surface could be used. The plates may have up to 384 wells, which allow us to check for many ligands in a relatively short time. Moreover, binding kinetics can be determined too.

C 4.5. CONCLUSIONS

i. Production of P34$_{pLS20}$FL protein turned out to be difficult since the protein seems to be unstable and we obtained a 55kDa-sized degradation of the protein, as proven by the PMF assay.

ii. P34$_{pLS20}$TED and P34$_{pLS20}$C68S structures have been solved. Structures reveal that P34$_{pLS20}$ is a TIE$_{pLS20}$ protein containing a Class II TED domain. P34$_{pLS20}$TED was obtained at a resolution of 1.57Å and belongs to the space group $P\ 4_32_12$, whereas P34$_{pLS20}$TEDC68S was obtained at a resolution of 2.49Å and belongs to the space group $P\ 4_12_12$.

iii. P34$_{pLS20}$TED structure is composed of five-stranded β-barrel and a three-helix bundle in the upper lobe and a 7 β-sandwich in the bottom lobe.

iv. A thioester bond is formed between cysteine 68 from β-strand C and glutamine 256 from β-strand Q.

v. Not all conserved motif in other Class II TED proteins are conserved in P34$_{pLS20}$TED. Cysteine 68 is located in the [YFL]CΦζ motif whereas glutamine 254 is located in the ΦQζΦΦ motif, as is the case in both Class I and Class II TEDs. However, the TQXXΦW motif found in other Class II TEDs is not preserved fully, as we can only find a tryptophan by superposition with other Class II TEDs, suggesting this motif is not essential.

vi. P34$_{pLS20}$TED and P34$_{pLS20}$C68S overall structures are virtually identical for the exception of the disruption of the thioester bond due to the entered mutation.

vii. P34$_{pLS20}$TED tertiary structure is highly similar to the already described other three Class II TED representatives BaTIE, SaTIE and EfmTIE86, having RMSDs of even 2.3Å

viii. Despite structure similarity, sequence identity is between 18 and 23%.

ix. The main difference between P34$_{pLS20}$TED and previously characterized class II TED representatives concern the specificity loop.

x. A further difference between P34$_{pLS20}$TED and BaTIE, SaTIE and EfmTIE86 lies in the anchor domain, as these three proteins have a LPXTG domain whereas P34$_{pLS20}$ is not predicted to have the same domain as a cell-anchor.

xi. The upper lobe of P34$_{pLS20}$TED corresponds to the canonical class I TED folding, while the bottom lobe corresponds to an Ig-like domain.

xii. Mutation of the cysteine 68 results in a loss of the thioester bond, as demonstrated both structurally, and in a loss of conjugation, as shown by functional studies.

xiii. $P34_{pLS20}$, and in particular C68S, has been demonstrated to be essential for the conjugative process.

xiv. Relevance of thioester bond has been shown to be remarkable as it is necessary for conjugation to happen.

xv. $P34_{pLS20}$ is the first structurally characterized mating pair formation protein encoded by a conjugative plasmid that contains a TED domain.

xvi. Crystallographic fragment screening methods were carried out to identify molecules that bind $P34_{pLS20}$TED. This led to huge amounts of data that we have not been able to completely process.

xvii. We suggest further experiments should be performed to study the binding partners of $P34_{pLS20}$TED.

BIBLIOGRAPHY

1. Segura, P. A., François, M., Gagnon, C. & Sauvé, S. Review of the occurrence of anti-infectives in contaminated wastewaters and natural and drinking waters. *Environmental Health Perspectives* (2009). doi:10.1289/ehp.11776

2. Roca, I. *et al.* The global threat of antimicrobial resistance: Science for intervention. *New Microbes and New Infections* (2015). doi:10.1016/j.nmni.2015.02.007

3. Cassini, A. *et al.* Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. *Lancet Infect. Dis.* (2019). doi:10.1016/S1473-3099(18)30605-4

4. Thorpe, K. E., Joski, P. & Johnston, K. J. Antibiotic-resistant infection treatment costs have doubled since 2002, now exceeding $2 billion annually. *Health Aff.* (2018). doi:10.1377/hlthaff.2017.1153

5. Bougnom, B. P. & Piddock, L. J. V. Wastewater for Urban Agriculture: A Significant Factor in Dissemination of Antibiotic Resistance. *Environmental Science and Technology* (2017). doi:10.1021/acs.est.7b01852

6. O'Neill, J. Antimicrobial Resistance : Tackling a crisis for the health and wealth of nations. *Rev. Antimicrob. Resist.* (2016).

7. WAKSMAN, S. A. What is an antibiotic or an antibiotic substance? *Mycologia* (1947). doi:10.2307/3755196

8. Mohr, K. I. History of antibiotics research. *Curr. Top. Microbiol. Immunol.* (2016). doi:10.1007/82_2016_499

9. Haensch, S. *et al.* Distinct clones of Yersinia pestis caused the black death. *PLoS Pathog.* (2010). doi:10.1371/journal.ppat.1001134

10. Kool, J. L. Risk of Person-to-Person Transmission of Pneumonic Plague. *Clin. Infect. Dis.* (2005). doi:10.1086/428617

11. Bassett, E. J., Keith, M. S., Armelagos, G. J., Martin, D. L. & Villanueva, A. R. Tetracycline-labeled human bone from ancient Sudanese Nubia (A.D. 350). *Science (80-.).* (1980). doi:10.1126/science.7001623

12. Nelson, M. L., Dinardo, A., Hochberg, J. & Armelagos, G. J. Brief communication: Mass spectroscopic characterization of tetracycline in the skeletal remains of an ancient population from Sudanese Nubia 350-550 CE. *Am. J. Phys. Anthropol.* (2010). doi:10.1002/ajpa.21340

13. Cook, M., Molto, E. & Anderson, C. Fluorochrome labelling in roman period skeletons from dakhleh oasis, Egypt. *Am. J. Phys. Anthropol.* (1989). doi:10.1002/ajpa.1330800202

14. Armelagos, G. J. Disease in ancient nubia. *Science (80-.).* (1969). doi:10.1126/science.163.3864.255

15. Joseph, I. O. F. *et al.* Proliferation of antibiotic-producing bacteria and concomitant antibiotic production as the basis for the antibiotic activity of Jordan's red soils. *Appl. Environ. Microbiol.* (2009). doi:10.1128/AEM.00104-09

16. Sobell, H. M. Actinomycin and DNA transcription. *Proc. Natl. Acad. Sci. U. S. A.* (1985). doi:10.1073/pnas.82.16.5328

17. Cui, L. & Su, X. Z. Discovery, mechanisms of action and combination therapy of artemisinin. *Expert Review of Anti-Infective Therapy* (2009). doi:10.1586/ERI.09.68

18. Wong, R. W. K. *et al.* Antimicrobial activity of Chinese medicine herbs against common bacteria in oral biofilm. A pilot study. *Int. J. Oral Maxillofac. Surg.* (2010). doi:10.1016/j.ijom.2010.02.024

19. Yuen, M. K. Z., Wong, R. W. K., Hägg, U. & Samaranayake, L. Antimicrobial Activity of Traditional Chinese Medicines on Common Oral Bacteria. *Chin. Med.* (2011). doi:10.4236/cm.2011.22007

20. Karamanou, M., Poulakou-Rebelakou, E., Tzetis, M. & Androutsos, G. Anton van Leeuwenhoek (1632-1723): Father of micromorphology and discoverer of spermatozoa. *Revista Argentina de Microbiologia* (2010). doi:10.1590/S0325-75412010000400013

21. Francis, E. Antony van Leeuwenhoek and his 'Little Animals'. *Science (80-.).* (1932). doi:10.1126/science.76.1982.597

22. Parkinson, J. *Theatrum botanicum = the theater of plants : or, An herball of a large extent ... /. Theatrum botanicum = the theater of plants : or, An herball of a large extent ... /* (2018). doi:10.5962/bhl.title.152383

23. Berche, P. Louis Pasteur, from crystals of life to vaccination. *Clinical Microbiology and Infection* (2012). doi:10.1111/j.1469-0691.2012.03945.x

24. Münch, R. Robert Koch. *Microbes and Infection* (2003). doi:10.1016/S1286-4579(02)00053-9

25. ABRAHAM, J. J. Some account of the history of the treatment of syphilis. *Br. J. Vener. Dis.* (1948). doi:10.1136/sti.24.4.153

26. Thomas, H. W. Some experiments in the treatment of trypanosomiasis. *Br. Med. J.* (1905). doi:10.1136/bmj.1.2317.1140

27. Ehrlich, P. & Hata, S. *Die experimentelle Chemotherapie der Spirillosen*. *Die experimentelle Chemotherapie der Spirillosen* (1910). doi:10.1007/978-3-642-64926-4

28. Williams, K. J. The introduction of 'chemotherapy' using arsphenamine - The first magic bullet. *Journal of the Royal Society of Medicine* (2009). doi:10.1258/jrsm.2009.09k036

29. Riethmiller, S. From atoxyl to salvarsan: Searching for the magic bullet. *Chemotherapy* (2005). doi:10.1159/000087453

30. Thorburn, A. L. Paul Ehrlich: pioneer of chemotherapy and cure by arsenic (1854-1915). *Br. J. Vener. Dis.* (1983). doi:10.1136/sti.59.6.404

31. Mahoney, J. F., Arnold, R. C. & Harris, A. Penicillin Treatment of Early Syphilis—A Preliminary Report. *Am. J. Public Heal. Nations Heal.* (1943). doi:10.2105/ajph.33.12.1387

32. Lloyd, N. C., Morgan, H. W., Nicholson, B. K. & Ronimus, R. S. The composition of Ehrlich's Salvarsan: Resolution of a century-old debate. *Angew. Chemie - Int. Ed.* (2005). doi:10.1002/anie.200461471

33. Lokaj, J. & John, C. Ilya Ilich Metchnikov and Paul Ehrlich: 1908 Nobel prize winners for their research on immunity. *Epidemiologie, Mikrobiologie, Imunologie* (2008).

34. Alexander Fleming. On the antibacterial action of cultures of a Penicillium, with special reference to their use in the isolation of B. injluenzae. *Br.]ournal Exp. Pathol.* (1929).

35. Chain, E. *et al.* PENICILLIN AS A CHEMOTHERAPEUTIC AGENT. *Lancet* (1940). doi:10.1016/S0140-6736(01)08728-1

36. Abraham, E. P. *et al.* FURTHER OBSERVATIONS ON PENICILLIN. *Lancet* (1941). doi:10.1016/S0140-6736(00)72122-2

37. Aldrich, S. Alexander Fleming Discovery and Development of Penicillin - Landmark - American Chemical Society. *American Chemical Society International Historic Chemical Landmarks* (1999). doi:https://doi.org/10.2307/3561468

38. Aminov, R. I. A brief history of the antibiotic era: Lessons learned and challenges for the future. *Front. Microbiol.* (2010). doi:10.3389/fmicb.2010.00134

39. Emmerich, R. & Löw, O. Bakteriolytische Enzyme als Ursache der erworbenen Immunität und die Heilung von Infectionskrankheiten durch dieselben. *Zeitschrift für Hyg. und Infect.* (1899). doi:10.1007/BF02206499

40. Bozdogan, B. & Appelbaum, P. C. Oxazolidinones: Activity, mode of action, and mechanism of resistance. *International Journal of Antimicrobial Agents* (2004). doi:10.1016/j.ijantimicag.2003.11.003

41. Munita, J. M., Arias, C. A., Unit, A. R. & Santiago, A. De. HHS Public Access Mechanisms of Antibiotic Resistance. *HHS Public Access* (2016). doi:10.1128/microbiolspec.VMBF-0016-2015.Mechanisms

42. Abraham, E. P. & Chain, E. An enzyme from bacteria able to destroy penicillin [1]. *Nature* (1940). doi:10.1038/146837a0

43. ROLLO, I. M., WILLIAMSON, J. & PLACKETT, R. L. Acquired resistance to penicillin and to neoarsphenamine in Spirochaeta recurrentis. *Br. J. Pharmacol. Chemother.* (1952). doi:10.1111/j.1476-5381.1952.tb00686.x

44. Cha, J. Y., Ishiwata, A. & Mobashery, S. A Novel β-Lactamase Activity from a Penicillin-binding Protein of Treponema pallidum and Why Syphilis Is Still Treatable with Penicillin. *J. Biol. Chem.* (2004). doi:10.1074/jbc.M400666200

45. Kumarasamy, K. K. *et al.* Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: A molecular, biological, and epidemiological study. *Lancet Infect. Dis.* (2010). doi:10.1016/S1473-3099(10)70143-2

46. Rammelkamp, C. H. & Maxon, T. Resistance of Staphylococcus aureus> to the Action of Penicillin. *Proc. Soc. Exp. Biol. Med.* (1942). doi:10.3181/00379727-51-13986

47. Lowy, F. D. Antimicrobial resistance: The example of Staphylococcus aureus. *Journal of Clinical Investigation* (2003). doi:10.1172/JCI18535

48. Hartman, B. & Tomasz, A. Altered penicillin-binding proteins in methicillin-resistant strains of Staphylococcus aureus. *Antimicrob. Agents Chemother.* (1981). doi:10.1128/AAC.19.5.726

49. Hansman, D., Devitt, L., Miles, H. & Riley, I. Pneumococci relatively insensitive to penicillin in Australia and New Guinea. *Med. J. Aust.* (1974). doi:10.5694/j.1326-5377.1974.tb70836.x

50. H.J., K., A., W. & K., K. Antimicrobial resistance in Streptococcus pneumoniae: A South African perspective. *Clin. Infect. Dis.* (1992).

51. Nordmann, P. Trends in β-Lactam Resistance Among Enterobacteriaceae. *Clin. Infect. Dis.* (1998). doi:10.1086/514905

52. Bouza, E. & Cercenado, E. Klebsiella and Enterobacter: Antibiotic resistance and treatment implications. *Seminars in Respiratory Infections* (2002). doi:10.1053/srin.2002.34693

53. Walsh, C. Molecular mechanisms that confer antibacterial drug resistance. *Nature* (2000). doi:10.1038/35021219

54. Lin, J. *et al.* Mechanisms of antibiotic resistance. *Frontiers in Microbiology* (2015). doi:10.3389/fmicb.2015.00034

55. Lobanovska, M. & Pilla, G. Penicillin's discovery and antibiotic resistance: Lessons for the future? *Yale J. Biol. Med.* (2017).

56. Tanwar, J., Das, S., Fatima, Z. & Hameed, S. Multidrug resistance: An emerging crisis. *Interdisciplinary Perspectives on Infectious Diseases* (2014). doi:10.1155/2014/541340

57. Fajardo, A. *et al.* The neglected intrinsic resistome of bacterial pathogens. *PLoS One* (2008). doi:10.1371/journal.pone.0001619

58. Munita, J. M. & Arias, C. A. Mechanisms of Antibiotic Resistance. in *Virulence Mechanisms of Bacterial Pathogens* (2016). doi:10.1128/9781555819286.ch17

59. Dantas, G. & Sommer, M. O. A. Context matters - the complex interplay between resistome genotypes and resistance phenotypes. *Current Opinion in Microbiology* (2012). doi:10.1016/j.mib.2012.07.004

60. Westra, E. R., Sünderhauf, D., Landsberger, M. & Buckling, A. Mechanisms and consequences of diversity-generating immune strategies. *Nature Reviews Immunology* (2017). doi:10.1038/nri.2017.78

61. Matic, I. *et al.* Highly variable mutation rates in commensal and pathogenic Escherichia coli. *Science (80-.).* (1997). doi:10.1126/science.277.5333.1833

62. Reams, A. B., Kofoid, E., Savageau, M. & Roth, J. R. Duplication frequency in a population of Salmonella enterica rapidly approaches steady state with or without recombination. *Genetics* (2010). doi:10.1534/genetics.109.111963

63. Darwin, C. *On the Origin of the Species*. *Darwin* (1859).

64. Goldenfeld, N. & Woese, C. Biology's next revolution. *Nature* (2007). doi:10.1038/445369a

65. Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. Mobile genetic elements: The agents of open source evolution. *Nature Reviews Microbiology* (2005). doi:10.1038/nrmicro1235

66. Durão, P., Balbontín, R. & Gordo, I. Evolutionary Mechanisms Shaping the Maintenance of Antibiotic Resistance. *Trends in Microbiology* (2018). doi:10.1016/j.tim.2018.01.005

67. Ochman, H., Lawrence, J. G. & Grolsman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* (2000). doi:10.1038/35012500

68. Lawrence, J. G. & Ochman, H. Molecular archaeology of the Escherichia coli genome. *Proc. Natl. Acad. Sci. U. S. A.* (1998). doi:10.1073/pnas.95.16.9413

69. Spellberg, B. *et al.* The Epidemic of Antibiotic-Resistant Infections: A Call to Action for the Medical Community from the Infectious Diseases Society of America. *Clin. Infect. Dis.* (2008). doi:10.1086/524891

70. Davies, J. Vicious circles: looking back on resistance plasmids. *Genetics* (1995).

71. Davies, J. Origins and evolution of antibiotic resistance. *Microbiología (Madrid, Spain)* (1996). doi:10.1128/mmbr.00016-10

72. Alekshun, M. N. & Levy, S. B. Molecular Mechanisms of Antibacterial Multidrug Resistance. *Cell* (2007). doi:10.1016/j.cell.2007.03.004

73. Wright, G. D. The antibiotic resistome: The nexus of chemical and genetic diversity. *Nature Reviews Microbiology* (2007). doi:10.1038/nrmicro1614

74. Cordero, O. X. & Hogeweg, P. The impact of long-distance horizontal gene transfer on prokaryotic genome size. *Proc. Natl. Acad. Sci. U. S. A.* (2009). doi:10.1073/pnas.0907584106

75. Griffith, F. The Significance of Pneumococcal Types. *J. Hyg. (Lond).* (1928). doi:10.1017/S0022172400031879

76. Avery, O. T., Macleod, C. M. & McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *J. Exp. Med.* (1944). doi:10.1084/jem.79.2.137

77. HOTCHKISS, R. D. Transfer of penicillin resistance in pneumococci by the desoxyribonucleate derived from resistant cultures. *Cold Spring Harb. Symp. Quant. Biol.* (1951). doi:10.1101/SQB.1951.016.01.032

78. ALEXANDER, H. E. & LEIDY, G. Induction of streptomycin resistance in sensitive Hemophilus influenzae by extracts containing desoxyribonucleic acid from resistant Hemophilus influenzae. *J. Exp. Med.* (1953). doi:10.1084/jem.97.1.17

79. ALEXANDER, H. E., HAHN, E. & LEIDY, G. On the specificity of the desoxyribonucleic acid which induces streptomycin resistance in Hemophilus. *J. Exp. Med.* (1956). doi:10.1084/jem.104.3.305

80. Zafra, O., Lamprecht-Grandío, M., de Figueras, C. G. & González-Pastor, J. E. Extracellular DNA Release by Undomesticated Bacillus subtilis Is Regulated by Early Competence. *PLoS One* (2012). doi:10.1371/journal.pone.0048716

81. Draghi, J. A. & Turner, P. E. DNA secretion and gene-level selection in bacteria. *Microbiology* (2006). doi:10.1099/mic.0.29013-0

82. Borgeaud, S., Metzger, L. C., Scrignari, T. & Blokesch, M. The type VI secretion system of Vibrio cholerae fosters horizontal gene transfer. *Science (80-.).* (2015). doi:10.1126/science.1260064

83. Chen, I., Christie, P. J. & Dubnau, D. The ins and outs of DNA transfer in bacteria. *Science* (2005). doi:10.1126/science.1114021

84. Chen, I. & Dubnau, D. DNA uptake during bacterial transformation. *Nature Reviews Microbiology* (2004). doi:10.1038/nrmicro844

85. Johnsborg, O. & Håvarstein, L. S. Regulation of natural genetic transformation and acquisition of transforming DNA in Streptococcus pneumoniae. in *FEMS Microbiology Reviews* (2009). doi:10.1111/j.1574-6976.2009.00167.x

86. Lorenz, M. G. & Wackernagel, W. Bacterial gene transfer by natural genetic transformation in the environment. *Microbiological Reviews* (1994).

87. Cohan, F. M., Roberts, M. S. & King, E. C. The Potential for Genetic Exchange by Transformation within a Natural Population of Bacillus subtilis. *Evolution (N. Y).* (1991). doi:10.2307/2409888

88. Johnston, C., Martin, B., Fichant, G., Polard, P. & Claverys, J. P. Bacterial transformation: Distribution, shared mechanisms and divergent control. *Nature Reviews Microbiology* (2014). doi:10.1038/nrmicro3199

89. Tsen, S. Der *et al.* Natural plasmid transformation in Escherichia coli. *J. Biomed. Sci.* (2002). doi:10.1159/000059425

90. Mirończuk, A. M., Kovács, Á. T. & Kuipers, O. P. Induction of natural competence in Bacillus cereus ATCC14579. *Microb. Biotechnol.* (2008). doi:10.1111/j.1751-7915.2008.00023.x

91. Meibom, K. L., Blokesch, M., Dolganov, N. A., Wu, C. Y. & Schoolnik, G. K. Microbiology: Chitin induces natural competence in vibrio cholerae. *Science (80-.).* (2005). doi:10.1126/science.1120096

92. Hamoen, L. W., Haijema, B., Bijlsma, J. J., Venema, G. & Lovett, C. M. The Bacillus subtilis Competence Transcription Factor, ComK, Overrides LexA-imposed Transcriptional Inhibition without Physically Displacing LexA. *J. Biol. Chem.* (2001). doi:10.1074/jbc.M104407200

93. Charpentier, X., Kay, E., Schneider, D. & Shuman, H. A. Antibiotics and UV radiation induce competence for natural transformation in Legionella pneumophila. *J. Bacteriol.* (2011). doi:10.1128/JB.01146-10

94. Prudhomme, M., Attaiech, L., Sanchez, G., Martin, B. & Claverys, J. P. Antibiotic stress induces genetic transformability in the human pathogen streptoccus pneumoniae. *Science (80-.).* (2006). doi:10.1126/science.1127912

95. Doolittle, W. F. *Lateral DNA transfer: Mechanisms and consequences*. *Nature* (2002). doi:10.1038/418589a

96. ZINDER, N. D. & LEDERBERG, J. Genetic exchange in Salmonella. *J. Bacteriol.* (1952).

97. Anthony JF Griffiths, Jeffrey H Miller, David T Suzuki, Richard C Lewontin, and W. M. G. *An Introduction to Genetic Analysis, 7th edition*. *ISBN* (2000).

98. Labrie, S. J., Samson, J. E. & Moineau, S. Bacteriophage resistance mechanisms. *Nature Reviews Microbiology* (2010). doi:10.1038/nrmicro2315

99. Mathews, C. K. Bacteriophage T4. in *eLS* (2015). doi:10.1002/9780470015902.a0000784.pub4

100. Molineux, I. J. The T7 Group. in *The Bacteriophages* (2006).

101. Casjens, S. R. & Hendrix, R. W. Bacteriophage lambda: Early pioneer and still relevant. *Virology* (2015). doi:10.1016/j.virol.2015.02.010

102. Tatum, E. L. & Lederberg, J. Gene Recombination in the Bacterium Escherichia coli. *J. Bacteriol.* (1947).

103. DAVIS, B. D. Nonfiltrability of the agents of genetic recombination in Escherichia coli. *J. Bacteriol.* (1950). doi:10.1128/jb.60.4.507-508.1950

104. Page, R. E., Gadau, J. & Beye, M. Perspectives Anecdotal, Historical and Critical Commentaries on Genetics The Emergence of Hymenopteran Genetics. *Genet. Soc. Am.* (2002).

105. Goessweiner-Mohr, N., Arends, K., Keller, W. & Grohmann, E. Conjugation in Gram-Positive Bacteria. *Microbiol. Spectr.* (2014). doi:10.1128/microbiolspec.plas-0004-2013

106. Johnson, C. M. & Grossman, A. D. Integrative and Conjugative Elements (ICEs): What They Do and How They Work. *Annu. Rev. Genet.* (2015). doi:10.1146/annurev-genet-112414-055018

107. Dubey, G. P. & Ben-Yehuda, S. Intercellular nanotubes mediate bacterial communication. *Cell* (2011). doi:10.1016/j.cell.2011.01.015

108. Mashburn-Warren, L. M. & Whiteley, M. Special delivery: Vesicle trafficking in prokaryotes. *Molecular Microbiology* (2006). doi:10.1111/j.1365-2958.2006.05272.x

109. Rosenshine, I., Tchelet, R. & Mevarech, M. The mechanism of DNA transfer in the mating system of an archaebacterium. *Science (80-.).* (1989). doi:10.1126/science.2818746

110. Lang, A. S., Zhaxybayeva, O. & Beatty, J. T. Gene transfer agents: Phage-like elements of genetic exchange. *Nature Reviews Microbiology* (2012). doi:10.1038/nrmicro2802

111. Earl, A. M., Losick, R. & Kolter, R. Ecology and genomics of Bacillus subtilis. *Trends in Microbiology* (2008). doi:10.1016/j.tim.2008.03.004

112. Henriques, A. O. & Moran, C. P. Structure, assembly, and function of the spore surface layers. *Annual Review of Microbiology* (2007). doi:10.1146/annurev.micro.61.080706.093224

113. Nicholson, W. L., Munakata, N., Horneck, G., Melosh, H. J. & Setlow, P. Resistance of Bacillus Endospores to Extreme Terrestrial and Extraterrestrial Environments. *Microbiol. Mol. Biol. Rev.* (2000). doi:10.1128/mmbr.64.3.548-572.2000

114. Hong, H. A. *et al.* Bacillus subtilis isolated from the human gastrointestinal tract. *Res. Microbiol.* (2009). doi:10.1016/j.resmic.2008.11.002

115. S., D. H. Beiträge zur Biologie der Pflanzen. *Nature* (1892). doi:10.1038/046461a0

116. Todar, K. Online Textbook of Bacteriology. *Bacterial Endotoxin* (2011).

117. Kunst, F. *et al.* The complete genome sequence of the gram-positive bacterium Bacillus subtilis. *Nature* (1997). doi:10.1038/36786

118. *Bacillus subtilis and Its Closest Relatives*. *Bacillus subtilis and Its Closest Relatives* (2002). doi:10.1128/9781555817992

119. Van Ert, M. N. *et al.* Global genetic population structure of Bacillus anthracis. *PLoS One* (2007). doi:10.1371/journal.pone.0000461

120. Cutting, S. M. Bacillus probiotics. *Food Microbiology* (2011). doi:10.1016/j.fm.2010.03.007

121. Monod, M., Denoya, C. & Dubnau, D. Sequence and properties of pIM13, a macrolide-lincosamide-streptogramin B resistance plasmid from Bacillus subtilis. *J. Bacteriol.* (1986). doi:10.1128/jb.167.1.138-147.1986

122. Phelan, R. W. *et al.* Tetracycline resistance-encoding plasmid from Bacillus sp. strain #24, isolated from the marine sponge haliclona simulans. *Appl. Environ. Microbiol.* (2011). doi:10.1128/AEM.01239-10

123. Roberts, A. P., Pratten, J., Wilson, M. & Mullany, P. Transfer of a conjugative transposon, Tn5397 in a model oral biofilm. *FEMS Microbiol. Lett.* (1999). doi:10.1016/S0378-1097(99)00288-8

124. Sommer, M. O. A., Dantas, G. & Church, G. M. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science (80-.).* (2009). doi:10.1126/science.1176950

125. Sankar, S. A., Lagier, J. C., Pontarotti, P., Raoult, D. & Fournier, P. E. The human gut microbiome, a taxonomic conundrum. *Systematic and Applied Microbiology* (2015). doi:10.1016/j.syapm.2015.03.004

126. Devirgiliis, C., Zinno, P. & Perozzi, G. Update on antibiotic resistance in foodborne Lactobacillus and Lactococcus species. *Front. Microbiol.* (2013). doi:10.3389/fmicb.2013.00301

127. Rankin, D. J., Rocha, E. P. C. & Brown, S. P. What traits are carried on mobile genetic elements, and why. *Heredity* (2011). doi:10.1038/hdy.2010.24

128. Stokes, H. W. & Gillings, M. R. Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into Gram-negative pathogens. *FEMS Microbiology Reviews* (2011). doi:10.1111/j.1574-6976.2011.00273.x

129. Leplae, R. ACLAME: A CLAssification of Mobile genetic Elements. *Nucleic Acids Res.* (2004). doi:10.1093/nar/gkh084

130. Hendrix, R. W. *et al.* Bacteriophages. in *Fields Virology: Sixth Edition* (2013). doi:10.5694/j.1326-5377.1923.tb118733.x

131. Sherratt, D. J. Bacterial plasmids. *Cell* (1974). doi:10.1016/0092-8674(74)90130-5

132. Curcio, M. J. & Derbyshire, K. M. The outs and ins of transposition: From MU to kangaroo. *Nature Reviews Molecular Cell Biology* (2003). doi:10.1038/nrm1241

133. Siguier, P., Gourbeyre, E. & Chandler, M. Bacterial insertion sequences: Their genomic impact and diversity. *FEMS Microbiol. Rev.* (2014). doi:10.1111/1574-6976.12067

134. Munoz-Lopez, M. & Garcia-Perez, J. DNA Transposons: Nature and Applications in Genomics. *Curr. Genomics* (2010). doi:10.2174/138920210790886871

135. D. Bryers, J. & R. Sharp, R. Comparison of retention and expression of recombinant plasmids between suspended and Biofilm-Bound bacteria degrading TCE. in *Progress in Biotechnology* (1996). doi:10.1016/S0921-0423(96)80033-5

136. Harvey, L. *et al. Molecular Cell Biology. 4th edition*. *Journal of the American Society for Mass Spectrometry* (2000). doi:10.1016/j.jasms.2009.08.001

137. del Solar, G., Giraldo, R., Ruiz-Echevarría, M. J., Espinosa, M. & Díaz-Orejas, R. Replication and control of circular bacterial plasmids. *Microbiol. Mol. Biol. Rev.* (1998).

138. Khan, S. A. Rolling-circle replication of bacterial plasmids. *Microbiol. Mol. Biol. Rev.* (1997).

139. Rodríguez, E. G. Nonviral DNA vectors for immunization and therapy: Design and methods for their obtention. *Journal of Molecular Medicine* (2004). doi:10.1007/s00109-004-0548-x

140. Ramsay, J. P. & Firth, N. Diverse mobilization strategies facilitate transfer of non-conjugative mobile genetic elements. *Current Opinion in Microbiology* (2017). doi:10.1016/j.mib.2017.03.003

141. Meijer, W. J. J., Venema, G. & Bron, S. Characterization of single strand origins of cryptic rolling-circle plasmids from Bacillus subtilis. *Nucleic Acids Res.* (1995). doi:10.1093/nar/23.4.612

142. Smillie, C., Garcillan-Barcia, M. P., Francia, M. V., Rocha, E. P. C. & de la Cruz, F. Mobility of Plasmids. *Microbiol. Mol. Biol. Rev.* (2010). doi:10.1128/mmbr.00020-10

143. Francia, M. V. *et al.* A classification scheme for mobilization regions of bacterial plasmids. *FEMS Microbiology Reviews* (2004). doi:10.1016/j.femsre.2003.09.001

144. Meijer, W. J. J. *et al.* Rolling-circle plasmids from Bacillus subtilis: Complete nucleolide sequences and analyses of genes of pTA1015, pTA1040, pTA1050 and pTA1060, and comparisons with related plasmids from Gram-positive bacteria. *FEMS Microbiol. Rev.* (1998). doi:10.1016/S0168-6445(98)00003-5

145. Meijer, W. J. J. *et al.* The endogenous Bacillus subtilis (natto) plasmids pTA1015 and pTA1040 contain signal peptidase-encoding genes: identification of a new structural module on cryptic plasmids. *Mol. Microbiol.* (1995). doi:10.1111/j.1365-2958.1995.mmi_17040621.x

146. Koehler, T. M. & Thorne, C. B. Bacillus subtilis (natto) plasmid pLS20 mediates interspecies plasmid transfer. *J. Bacteriol.* (1987). doi:10.1128/jb.169.11.5271-5278.1987

147. Paulsson, J. Multileveled selection on plasmid replication. *Genetics* (2002).

148. Kubo, Y. *et al.* Phylogenetic analysis of Bacillus subtilis strains applicable to natto (fermented soybean) production. *Appl. Environ. Microbiol.* (2011). doi:10.1128/AEM.00448-11

149. Hara, T., Aumayr, A., Fujio, Y. & Ueda, S. Elimination of plasmid-linked polyglutamate production by Bacillus subtilis (natto) with acridine orange. *Appl. Environ. Microbiol.* (1982). doi:10.1128/aem.44.6.1456-1458.1982

150. Hara, T., Ueda, S. & Aumayr, A. Characterization of plasmid deoxyri-bonucleic acid in Bacillus Natto: Evidence for plasmid-linked PGA production. *J. Gen. Appl. Microbiol.* (1981). doi:10.2323/jgam.27.299

151. Tanaka, T. & Koshikawa, T. Isolation and characterization of four types of plasmids from Bacillus subtilis (natto). *J. Bacteriol.* (1977).

152. Bauer, T., Rösch, T., Itaya, M. & Graumann, P. L. Localization pattern of conjugation machinery in a Gram-positive bacterium. *J. Bacteriol.* (2011). doi:10.1128/JB.00175-11

153. Singh, P. K. *et al.* Inhibition of Bacillus subtilis natural competence by a native, conjugative plasmid-encoded comK repressor protein. *Environ. Microbiol.* (2012). doi:10.1111/j.1462-2920.2012.02819.x

154. Miyano, M. *et al.* Rapid conjugative mobilization of a 100kb segment of Bacillus subtilis chromosomal DNA is mediated by a helper plasmid with no ability for self-transfer. *Microb. Cell Fact.* (2018). doi:10.1186/s12934-017-0855-x

155. Meijer, W. J. J., Boer, A. J. D., Tongeren, S. Van, Venema, G. & Bron, S. Characterization of the replication region of the acillus subtilis plasmid pLs20: A noval type to replicon. *Nucleic Acids Res.* (1995). doi:10.1093/nar/23.16.3214

156. Cuatrecasas, P., Wilchek, M. & Anfinsen, C. B. Selective enzyme purification by affinity chromatography. *Proc. Natl. Acad. Sci. U. S. A.* (1968). doi:10.1073/pnas.61.2.636

157. Porath, J. From gel filtration to adsorptive size exclusion. in *Journal of Protein Chemistry* (1997). doi:10.1023/A:1026357326667

158. Coskun, O. Separation Tecniques: CHROMATOGRAPHY. *North. Clin. Istanbul* (2016). doi:10.14744/nci.2016.32757

159. In, M. Guide to Protein Purification, 2nd Edition. *Methods Enzymol.* (2009). doi:10.1016/S0076-6879(09)63045-7

160. Lowe, C. R. Combinatorial approaches to affinity chromatography. *Current Opinion in Chemical Biology* (2001). doi:10.1016/S1367-5931(00)00199-X

161. Smyth, M. S. & Martin, J. H. J. x Ray crystallography. *Journal of Clinical Pathology - Molecular Pathology* (2000). doi:10.1136/mp.53.1.8

162. Hendrickson, W. A. Anomalous diffraction in crystallographic phase evaluation. *Quarterly Reviews of Biophysics* (2014). doi:10.1017/S0033583514000018

163. Argos, P. & Rossman, M. G. Molecular Replacement Method. in *Theory and Practice of Direct Methods in Crystallography* (1980). doi:10.1007/978-1-4613-2979-4_10

164. Evans, P. & McCoy, A. An introduction to molecular replacement. in *Acta Crystallographica Section D: Biological Crystallography* (2007). doi:10.1107/S0907444907051554

165. Giacovazzo, C. Direct methods. in *International Tables for Crystallography* (2006). doi:10.1107/97809553602060000555

166. Rodríguez, D. D. *et al.* Crystallographic ab initio protein structure solution below atomic resolution. *Nat. Methods* (2009). doi:10.1038/nmeth.1365

167. Rodríguez, D. *et al.* Practical structure solution with ARCIMBOLDO. *Acta Crystallogr. Sect. D Biol. Crystallogr.* (2012). doi:10.1107/S0907444911056071

168. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* (2007). doi:10.1107/S0021889807021206

169. Sheldrick, G. M. Macromolecular phasing with SHELXE. in *Zeitschrift fur Kristallographie* (2002). doi:10.1524/zkri.217.12.644.20662

170. Luscombe, N. M., Greenbaum, D. & Gerstein, M. What is bioinformatics? An introduction and overview. *Yearb. Med. Inform.* (2001). doi:10.1055/s-0038-1638103

171. Madeira, F. *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gkz268

172. Clamp, M., Cuff, J., Searle, S. M. & Barton, G. J. The Jalview Java alignment editor. *Bioinformatics* (2004). doi:10.1093/bioinformatics/btg430

173. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* (1990).

doi:10.1016/S0022-2836(05)80360-2

174. Krissinel, E. & Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. Sect. D Biol. Crystallogr.* (2004). doi:10.1107/S0907444904026460

175. Holm, L. DALI and the persistence of protein shape. *Protein Sci.* (2020). doi:10.1002/pro.3749

176. Christie, P. J., Atmakuri, K., Krishnamoorthy, V., Jakubowski, S. & Cascales, E. BIOGENESIS, ARCHITECTURE, AND FUNCTION OF BACTERIAL TYPE IV SECRETION SYSTEMS. *Annu. Rev. Microbiol.* (2005). doi:10.1146/annurev.micro.58.030603.123630

177. Korobkova, E., Emonet, T., Vilar, J. M. G., Shimizu, T. S. & Cluzel, P. From molecular noise to behavioural variability in a single bacterium. *Nature* (2004). doi:10.1038/nature02404

178. Dubnau, D. & Losick, R. Bistability in bacteria. *Molecular Microbiology* (2006). doi:10.1111/j.1365-2958.2006.05249.x

179. Veening, J.-W., Smits, W. K. & Kuipers, O. P. Bistability, Epigenetics, and Bet-Hedging in Bacteria. *Annu. Rev. Microbiol.* (2008). doi:10.1146/annurev.micro.62.081307.163002

180. Ramachandran, G. *et al.* A Complex Genetic Switch Involving Overlapping Divergent Promoters and DNA Looping Regulates Expression of Conjugation Genes of a Gram-positive Plasmid. *PLoS Genet.* (2014). doi:10.1371/journal.pgen.1004733

181. Singh, P. K. *et al.* Mobility of the Native Bacillus subtilis Conjugative Plasmid pLS20 Is Regulated by Intercellular Signaling. *PLoS Genet.* (2013). doi:10.1371/journal.pgen.1003892

182. Bendtsen, K. M. *et al.* Direct and indirect effects in the regulation of overlapping promoters. *Nucleic Acids Res.* (2011). doi:10.1093/nar/gkr390

183. KANIA, J. & MÜLLER-HILL, B. Construction, Isolation and Implications of Repressor-galactosidase ·ß-galactosidase Hybrid Molecules. *Eur. J. Biochem.* (1977). doi:10.1111/j.1432-1033.1977.tb11819.x

184. Dunn, T. M., Hahn, S., Ogden, S. & Schleif, R. F. An operator at -280 base pairs that is required for repression of araBAD operon promoter: Addition of DNA helical turns between the operator and promoter cyclically hinders repression. *Proc. Natl. Acad. Sci. U. S. A.* (1984). doi:10.1073/pnas.81.16.5017

185. Matthews, K. S. DNA looping. *Microbiological Reviews* (1992). doi:10.1146/annurev.biochem.61.1.199

186. Mukherjee, S., Erickson, H. & Bastia, D. Enhancer-origin interaction in plasmid R6K involves a DNA loop mediated by initiator protein. *Cell* (1988). doi:10.1016/S0092-8674(88)80030-8

187. Bondarenko, V. A., Liu, Y. V., Jiang, Y. I. & Studitsky, V. M. Communication over a large distance: Enhancers and insulators. in *Biochemistry and Cell Biology* (2003). doi:10.1139/o03-051

188. Oehler, S. & Müller-Hill, B. High Local Concentration: A Fundamental Strategy of Life. *Journal of Molecular Biology* (2010). doi:10.1016/j.jmb.2009.10.056

189. Cournac, A. & Plumbridge, J. DNA Looping in Prokaryotes: Experimental and theoretical approaches. *J. Bacteriol.* (2013). doi:10.1128/JB.02038-12

190. Rao, C. V., Wolf, D. M. & Arkin, A. P. Control, exploitation and tolerance of intracellular noise. *Nature* (2002). doi:10.1038/nature01258

191. Wang, Z. & Zhang, J. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *Proc. Natl. Acad. Sci. U. S. A.* (2011). doi:10.1073/pnas.1100059108

192. Pottathil, M. & Lazazzera, B. A. The extracellular PHR peptide-rap phosphatase signaling circuit of bacillus subtilis. *Frontiers in Bioscience* (2003). doi:10.2741/913

193. Koetje, E. J., Hajdo-Milasinovic, A., Kiewiet, R., Bron, S. & Tjalsma, H. A plasmid-borne Rap-Phr system of Bacillus subtilis can mediate cell-density controlled production of extracellular proteases. *Microbiology* (2003). doi:10.1099/mic.0.25737-0

194. Schultz, D., Wolynes, P. G., Jacob, E. Ben & Onuchic, J. N. Deciding fate in adverse times: Sporulation and competence in Bacillus subtilis. *Proc. Natl. Acad. Sci. U. S. A.* (2009). doi:10.1073/pnas.0912185106

195. Crespo, I. *et al.* Inactivation of the dimeric RappLS20 anti-repressor of the conjugation operon is mediated by peptide-induced tetramerization. *Nucleic Acids Res.* (2020). doi:10.1093/nar/gkaa540

196. Juanhuix, J. *et al.* Developments in optics and performance at BL13-XALOC, the macromolecular crystallography beamline at the Alba Synchrotron. *J. Synchrotron Radiat.* (2014). doi:10.1107/S160057751400825X

197. Vonrhein, C. *et al.* Data processing and analysis with the {\it autoPROC} toolbox. *Acta Crystallogr. Sect. D* **67**, 293–302 (2011).

198. Tickle, I. J. *et al.* STARANISO. (2018).

199. Adams, P. D. *et al.* PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* (2010). doi:10.1107/S0907444909052925

200. DeLano, W. L. Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr.* (2002).

201. The CCP4 suite: Programs for protein crystallography. *Acta Crystallogr. Sect. D Biol. Crystallogr.* (1994). doi:10.1107/S0907444994003112

202. Berman, H. M. *et al.* The protein data bank. *Acta Crystallogr. Sect. D Biol. Crystallogr.* (2002). doi:10.1107/S0907444902003451

203. Davison, T. S. *et al.* Structure and functionality of a designed p53 dimer. *J. Mol. Biol.* (2001). doi:10.1006/jmbi.2001.4450

204. Pietsch, E. C., Sykes, S. M., McMahon, S. B. & Murphy, M. E. The p53 family and programmed cell death. *Oncogene* (2008). doi:10.1038/onc.2008.315

205. Yang, A., Kaghad, M., Caput, D. & McKeon, F. On the shoulders of giants: p63, p73 and the rise of p53. *Trends in Genetics* (2002). doi:10.1016/S0168-9525(02)02595-7

206. Levrero, M. *et al.* The p53/p63/p73 family of transcription factors: Overlapping and distinct functions. *Journal of Cell Science* (2000).

207. Riley, T., Sontag, E., Chen, P. & Levine, A. Transcriptional control of human p53-regulated genes. *Nat. Rev. Mol. Cell Biol.* (2008). doi:10.1038/nrm2395

208. Costanzo, A. *et al.* DNA damage-dependent acetylation of p73 dictates the selective activation of apoptotic target genes. *Mol. Cell* (2002). doi:10.1016/S1097-2765(02)00431-8

209. Petitjean, A. *et al.* Properties of the six isoforms of p63: P53-like regulation in response to genotoxic stress and cross talk with δNp73. *Carcinogenesis* (2008). doi:10.1093/carcin/bgm258

210. Fatt, M. P., Cancino, G. I., Miller, F. D. & Kaplan, D. R. P63 and p73 coordinate p53 function to determine the balance between survival, cell death, and senescence in adult neural precursor cells. *Cell Death Differ.* (2014). doi:10.1038/cdd.2014.61

211. Flores, E. R. *et al.* p63 and p73 are required for p53-dependent apoptosis in response to DNA damage. *Nature* (2002). doi:10.1038/416560a

212. Vogelstein, B., Lane, D. & Levine, A. J. Surfing the p53 network. *Nature* (2000). doi:10.1038/35042675

213. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* (2013). doi:10.1038/nature12634

214. Ball, V. A. *et al.* p53 immunostaining of surgical margins as a predictor of local recurrence in squamous cell carcinoma of the oral cavity and oropharynx. *Ear, Nose Throat J.* (1997). doi:10.1177/014556139707601109

215. Petitjean, A. *et al.* Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: Lessons from recent developments in the IARC TP53 database. *Hum. Mutat.* (2007). doi:10.1002/humu.20495

216. Yang, A. *et al.* p73-Deficient mice have neurological, pheromonal and inflammatory defects but lack spontaneous tumours. *Nature* (2000). doi:10.1038/35003607

217. Yang, A. *et al.* p63, a p53 homolog at 3q27-29, encodes multiple products with transactivating, death-inducing, and dominant-negative activities. *Mol. Cell* (1998). doi:10.1016/S1097-2765(00)80275-0

218. Suh, E. K. *et al.* p63 protects the female germ line during meiotic arrest. *Nature* (2006). doi:10.1038/nature05337

219. Itahana, Y., Ke, H. & Zhang, Y. p53 oligomerization is essential for its C-terminal lysine acetylation. *J. Biol. Chem.* (2009). doi:10.1074/jbc.M805696200

220. Kitayner, M. *et al.* Structural Basis of DNA Recognition by p53 Tetramers. *Mol. Cell* (2006). doi:10.1016/j.molcel.2006.05.015

221. Tidow, H. *et al.* Quaternary structures of tumor suppressor p53 and a specific p53-DNA complex. *Proc. Natl. Acad. Sci. U. S. A.* (2007). doi:10.1073/pnas.0705069104

222. Laptenko, O., Tong, D. R., Manfredi, J. & Prives, C. The Tail That Wags the Dog: How the Disordered C-Terminal Domain Controls the Transcriptional Activities of the p53 Tumor-Suppressor Protein. *Trends in Biochemical Sciences* (2016). doi:10.1016/j.tibs.2016.08.011

223. Schultz, J., Bork, P., Ponting, C. P. & Hofmann, K. SAM as a protein interaction domain involved in developmental regulation. *Protein Sci.* (2008). doi:10.1002/pro.5560060128

224. Scoumanne, A., Harms, K. L. & Chen, X. Structural basis for gene activation by p53 family members. *Cancer Biology and Therapy* (2005). doi:10.4161/cbt.4.11.2254

225. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* (1999). doi:10.1006/jmbi.1999.3091

226. Buchan, D. W. A. & Jones, D. T. The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gkz297

227. Brodsky, M. H. *et al.* Comparison of this sequence with the database of Dro. *Cell* (2000).

228. Huyen, Y. *et al.* Structural differences in the DNA binding domains of human p53 and its C. elegans ortholog Cep-1. *Structure* (2004). doi:10.1016/j.str.2004.05.007

229. Jeffrey, P. D., Gorina, S. & Pavletich, N. P. Crystal structure of the tetramerization domain of the p53 tumor suppressor at 1.7 angstroms. *Science (80-.).* (1995). doi:10.1126/science.7878469

230. Ou, H. Der, Löhr, F., Vogel, V., Mäntele, W. & Dötsch, V. Structural evolution of C-terminal domains in the p53 family. *EMBO J.* (2007). doi:10.1038/sj.emboj.7601764

231. Joerger, A. C. & Fersht, A. R. Structural Biology of the Tumor Suppressor p53. *Annu. Rev. Biochem.* (2008). doi:10.1146/annurev.biochem.77.060806.091238

232. Zaika, A. I., Wei, J., Noto, J. M. & Peek, R. M. Microbial Regulation of p53 Tumor Suppressor. *PLoS Pathogens* (2015). doi:10.1371/journal.ppat.1005099

233. Chang, A. H. & Parsonnet, J. Role of bacteria in oncogenesis. *Clinical Microbiology Reviews* (2010). doi:10.1128/CMR.00012-10

234. Thoendel, M. & Horswill, A. R. Biosynthesis of peptide signals in gram-positive bacteria. *Advances in applied microbiology* (2010). doi:10.1016/S0065-2164(10)71004-2

235. Waters, C. M. & Bassler, B. L. QUORUM SENSING: Cell-to-Cell Communication in Bacteria. *Annu. Rev. Cell Dev. Biol.* (2005). doi:10.1146/annurev.cellbio.21.012704.131001

236. Rocha-Estrada, J., Aceves-Diez, A. E., Guarneros, G. & De La Torre, M. The RNPP family of quorum-sensing proteins in Gram-positive bacteria. *Applied Microbiology and Biotechnology* (2010). doi:10.1007/s00253-010-2651-y

237. Neiditch, M. B., Capodagli, G. C., Prehna, G. & Federle, M. J. Genetic and Structural Analyses of RRNPP Intercellular Peptide Signaling of Gram-Positive Bacteria. *Annu. Rev. Genet.* (2017). doi:10.1146/annurev-genet-120116-023507

238. Bareia, T., Pollak, S. & Eldar, A. Self-sensing in Bacillus subtilis quorum-sensing systems. *Nat. Microbiol.* (2018). doi:10.1038/s41564-017-0044-z

239. Dunny, G. M. & Leonard, B. A. B. CELL-CELL COMMUNICATION IN GRAM-POSITIVE BACTERIA. *Annu. Rev. Microbiol.* (1997). doi:10.1146/annurev.micro.51.1.527

240. Kohler, V., Keller, W. & Grohmann, E. Regulation of gram-positive conjugation. *Frontiers in Microbiology* (2019). doi:10.3389/fmicb.2019.01134

241. Ohara, M., Wu, H. C., Sankaran, K. & Rick, P. D. Identification and characterization of a new lipoprotein, NlpI, in Escherichia coli K-12. *J. Bacteriol.* (1999). doi:10.1128/jb.181.14.4318-4325.1999

242. Gallego del Sol, F., Penadés, J. R. & Marina, A. Deciphering the Molecular Mechanism Underpinning Phage Arbitrium Communication Systems. *Mol. Cell* (2019). doi:10.1016/j.molcel.2019.01.025

243. Parashar, V., Mirouze, N., Dubnau, D. A. & Neiditch, M. B. Structural basis of response regulator dephosphorylation by rap phosphatases. *PLoS Biol.* (2011). doi:10.1371/journal.pbio.1000589

244. Perego, M. *et al.* Multiple protein-aspartate phosphatases provide a mechanism for the integration of diverse signals in the control of development in B. subtilis. *Cell* (1994). doi:10.1016/0092-8674(94)90035-3

245. Meijer, W. J. J. *et al.* Rolling-circle plasmids from Bacillus subtilis : complete nucleotide sequences and analyses of genes of pTA1015, pTA1040, pTA1050 and pTA1060, and comparisons with related plasmids from Gram-positive bacteria . *FEMS Microbiol. Rev.* (1998). doi:10.1111/j.1574-6976.1998.tb00357.x

246. Zhu, J. *et al.* LGN/mInsc and LGN/NuMA Complex Structures Suggest Distinct Functions in Asymmetric Cell Division for the Par3/mInsc/LGN and Gαi/LGN/NuMA Pathways. *Mol. Cell* (2011). doi:10.1016/j.molcel.2011.07.011

247. Baker, M. D. & Neiditch, M. B. Structural basis of response regulator inhibition by a bacterial anti-activator protein. *PLoS Biol.* (2011). doi:10.1371/journal.pbio.1001226

248. Parashar, V., Jeffrey, P. D. & Neiditch, M. B. Conformational Change-Induced Repeat Domain Expansion Regulates Rap Phosphatase Quorum-Sensing Signal Receptors. *PLoS Biol.* (2013). doi:10.1371/journal.pbio.1001512

249. Declerck, N. *et al.* Structure of PlcR: Insights into virulence regulation and evolution of quorum sensing in Gram-positive bacteria. *Proc. Natl. Acad. Sci. U. S. A.* (2007). doi:10.1073/pnas.0704501104

250. Bae, T. & Dunny, G. M. Dominant-negative mutants of prgX: Evidence for a role for PrgX dimerization in negative regulation of pheromone-inducible conjugation. *Mol. Microbiol.* (2001). doi:10.1046/j.1365-2958.2001.02319.x

251. Kozlowicz, B. K. *et al.* Molecular basis for control of conjugation by bacterial pheromone and inhibitor peptides. *Mol. Microbiol.* (2006). doi:10.1111/j.1365-2958.2006.05434.x

252. Zouhir, S. *et al.* Peptide-binding dependent conformational changes regulate the transcriptional activity of the quorum-sensor NprR. *Nucleic Acids Res.* (2013). doi:10.1093/nar/gkt546

253. Crespo, I. *et al.* Inactivation of the dimeric RappLS20 anti-repressor of the conjugation operon is mediated by peptide-induced tetramerization. *Nucleic Acids Res.* (2020). doi:10.1093/nar/gkaa540

254. Rösch, T. C. & Graumann, P. L. Induction of plasmid conjugation in Bacillus subtilis is bistable and driven by a direct interaction of a Rap/Phr quorum-sensing system with a master repressor. *J. Biol. Chem.* (2015). doi:10.1074/jbc.M115.664110

255. Core, L. & Perego, M. TPR-mediated interaction of RapC with ComA inhibits response regulator-DNA binding for competence development in Bacillus subtilis. *Mol. Microbiol.* (2003). doi:10.1046/j.1365-2958.2003.03659.x

256. Bongiorni, C., Ishikawa, S., Stephenson, S., Ogasawara, N. & Perego, M. Synergistic regulation of competence development in Bacillus subtilis by two Rap-Phr systems. *J. Bacteriol.* (2005). doi:10.1128/JB.187.13.4353-4361.2005

257. Vasu, K. & Nagaraja, V. Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense. *Microbiol. Mol. Biol. Rev.* (2013). doi:10.1128/mmbr.00044-12

258. Vasu, K. & Nagaraja, V. Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense. *Microbiol. Mol. Biol. Rev.* (2013). doi:10.1128/mmbr.00044-12

259. Jones, A. L., Barth, P. T. & Wilkins, B. M. Zygotic induction of plasmid ssb and psiB genes following conjugative transfer of Incl1 plasmid Collb-P9. *Mol. Microbiol.* (1992). doi:10.1111/j.1365-2958.1992.tb01507.x

260. Petrova, V., Chitteni-Pattu, S., Drees, J. C., Inman, R. B. & Cox, M. M. An SOS Inhibitor that Binds to Free RecA Protein: The PsiB Protein. *Mol. Cell* (2009). doi:10.1016/j.molcel.2009.07.026

261. Walker, G. C. The SOS Response of Escherichia coli. *Compr. Rev.* (1987).

262. Golub, E., Bailone, A. & Devoret, R. A gene encoding an SOS inhibitor is present in different conjugative plasmids. *J. Bacteriol.* (1988). doi:10.1128/jb.170.9.4392-4394.1988

263. Chilley, P. M. & Wilkins, B. M. Distribution of the ardA family of antirestriction genes on conjugative plasmids. *Microbiology* (1995). doi:10.1099/13500872-141-9-2157

264. Masai, H. & Arai, K. I. Frpo: A novel single-stranded DNA promoter for transcription and for primer RNA synthesis of DNA replication. *Cell* (1997). doi:10.1016/S0092-8674(00)80275-5

265. Althorpe, N. J., Chilley, P. M., Thomas, A. T., Brammar, W. J. & Wilkins, B. M. Transient transcriptional activation of the Incl1 plasmid anti-restriction gene (ardA) and SOS inhibition gene (psiB) early in conjugating recipient. *Mol. Microbiol.* (1999). doi:10.1046/j.1365-2958.1999.01153.x

266. Bagdasarian, M. *et al.* PsiB, an anti-SOS protein, is transiently expressed by the F sex factor during its transmission to an Escherichia coli K-12 recipient. *Mol. Microbiol.* (1992). doi:10.1111/j.1365-2958.1992.tb01539.x

267. Singh, P. K., Ballestero-Beltrán, S., Ramachandran, G. & Meijer, W. J. J. Complete nucleotide sequence and determination of

the replication region of the sporulation inhibiting plasmid p576 from Bacillus pumilus NRS576. *Res. Microbiol.* (2010). doi:10.1016/j.resmic.2010.07.007

268. Val-Calvo, J. *et al.* Novel regulatory mechanism of establishment genes of conjugative plasmids. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gky996

269. Harrison, S. DNA Recognition By Proteins With The Helix-Turn-Helix Motif. *Annu. Rev. Biochem.* (1990). doi:10.1146/annurev.biochem.59.1.933

270. Knight, K. L. & Sauer, R. T. DNA binding specificity of the Arc and Mnt repressors is determined by a short region of N-terminal residues. *Proc. Natl. Acad. Sci. U. S. A.* (1989). doi:10.1073/pnas.86.3.797

271. Knight, K. L., Bowie, J. U., Vershon, A. K., Kelley, R. D. & Sauer, R. T. The Arc and Mnt repressors. A new class of sequence-specific DNA-binding protein. *J. Biol. Chem.* (1989).

272. Breg, J. N., Van Opheusden, J. H. J., Burgering, M. J. M., Boelens, R. & Kaptein, R. Structure of Arc represser in solution: evidence for a family of β-sheet DMA-binding proteins. *Nature* (1990). doi:10.1038/346586a0

273. Somers, W. S. & Phillips, S. E. V. Crystal structure of the met represser-operator complex at 2.8 Å resolution reveals DNA recognition by β-strands. *Nature* (1992). doi:10.1038/359387a0

274. Raumann, B. E., Rould, M. A., Pabo, C. O. & Sauer, R. T. DNA recognition by β-sheets in the Arc represser-operator crystal structure. *Nature* (1994). doi:10.1038/367754a0

275. Schreiter, E. R. & Drennan, C. L. Ribbon-helix-helix transcription factors: Variations on a theme. *Nat. Rev. Microbiol.* (2007). doi:10.1038/nrmicro1717

276. Miguel-Arribas, A. *et al.* The Bacillus subtilis conjugative plasmid pLS20 encodes two ribbon-helix-helix type auxiliary relaxosome proteins that are essential for conjugation. *Front. Microbiol.* (2017). doi:10.3389/fmicb.2017.02138

277. Yoshida, H. *et al.* Structural Basis of the Role of the NikA Ribbon-Helix-Helix Domain in Initiating Bacterial Conjugation. *J. Mol. Biol.* (2008). doi:10.1016/j.jmb.2008.09.067

278. Wong, J. J. W., Lu, J., Edwards, R. A., Frost, L. S. & Glover, J. N. M. Structural basis of cooperative DNA recognition by the plasmid conjugation factor, TraM. *Nucleic Acids Res.* (2011). doi:10.1093/nar/gkr296

279. Lujan, S. A., Guogas, L. M., Ragonese, H., Matson, S. W. & Redinbo, M. R. Disrupting antibiotic resistance propagation by inhibiting the conjugative DNA relaxase. *Proc. Natl. Acad. Sci. U. S. A.* (2007). doi:10.1073/pnas.0702760104

280. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D: Biological Crystallography* (2011). doi:10.1107/S0907444910045749

281. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. Sect. D Biol. Crystallogr.* (2012). doi:10.1107/S0907444912001308

282. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. Sect. D Biol. Crystallogr.* (2010). doi:10.1107/S0907444910007493

283. Brown, B. M. & Sauer, R. T. Assembly of the Arc Repressor-Operator Complex: Cooperative Interactions between DNA-Bound Dimers. *Biochemistry* (1993). doi:10.1021/bi00056a022

284. Pizarro-Cerdá, J. & Cossart, P. Bacterial adhesion and entry into host cells. *Cell* (2006). doi:10.1016/j.cell.2006.02.012

285. Linke-Winnebeck, C. *et al.* Structural model for covalent adhesion of the Streptococcus pyogenes pilus through a thioester bond. *J. Biol. Chem.* (2014). doi:10.1074/jbc.M113.523761

286. Pointon, J. A. *et al.* A highly unusual thioester bond in a pilus adhesin is required for efficient host cell interaction. *J. Biol. Chem.* (2010). doi:10.1074/jbc.M110.149385

287. Schwarz-Linek, U. & Banfield, M. J. Yet more intramolecular cross-links in Gram-positive surface proteins. *Proceedings of the National Academy of Sciences of the United States of America* (2014). doi:10.1073/pnas.1322482111

288. Hae, J. K., Coulibaly, F., Clow, F., Proft, T. & Baker, E. N. Stabilizing isopeptide bonds revealed in gram-positive bacterial pilus structure. *Science (80-.).* (2007). doi:10.1126/science.1145806

289. Kwon, H., Squire, C. J., Young, P. G. & Baker, E. N. Autocatalytically generated Thr-Gln ester bond cross-links stabilize the repetitive Ig-domain shaft of a bacterial cell surface adhesin. *Proc. Natl. Acad. Sci. U. S. A.* (2014). doi:10.1073/pnas.1316855111

290. Walden, M. *et al.* An internal thioester in a pathogen surface protein mediates covalent host binding. *Elife* (2015). doi:10.7554/eLife.06638

291. Miller, O. K., Banfield, M. J. & Schwarz-Linek, U. A new structural class of bacterial thioester domains reveals a slipknot topology. *Protein Sci.* (2018). doi:10.1002/pro.3478

292. Echelman, D. J., Lee, A. Q. & Fernández, J. M. Mechanical forces regulate the reactivity of a thioester bond in a bacterial adhesin. *Journal of Biological Chemistry* (2017). doi:10.1074/jbc.M117.777466

293. Dodds, A. W. & Law, S. K. A. The phylogeny and evolution of the thioester bond-containing proteins C3, C4 and α2-macroglobulin. *Immunological Reviews* (1998). doi:10.1111/j.1600-065X.1998.tb01249.x

294. Cherry, S. & Silverman, N. Host-pathogen interactions in drosophila: New tricks from an old friend. *Nature Immunology* (2006). doi:10.1038/ni1388

295. Baker, E. N. & Young, P. G. Convergent weaponry in a biological arms race. *Elife* (2015). doi:10.7554/eLife.08710

296. Janssen, B. J. C., Christodoulidou, A., McCarthy, A., Lambris, J. D. & Gros, P. Structure of C3b reveals conformational changes that underlie complement activity. *Nature* (2006). doi:10.1038/nature05172

297. Law, S. K. A. & Dodds, A. W. The internal thioester and the covalent binding properties of the complement proteins C3 and C4. *Protein Sci.* (2008). doi:10.1002/pro.5560060201

298. Kabsch, W. *et al. XDS*. *Acta Crystallogr. Sect. D Biol. Crystallogr.* (2010). doi:10.1107/S0907444909047337

299. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. Sect. D Biol. Crystallogr.* (2013). doi:10.1107/S0907444913000061

300. Perrakis, A., Morris, R. & Lamzin, V. S. Automated protein model building combined with iterative structure refinement. *Nat. Struct. Biol.* (1999). doi:10.1038/8263

301. DeLano, W. L. The PyMOL Molecular Graphics System, Version 2.3. *Schrödinger LLC* (2020). doi:10.1038/hr.2014.17

302. Mattos, C. & Ringe, D. Locating and Characterizing Binding Sites on Proteins. *Nat. Biotechnol.* (1996). doi:10.1038/nbt0596-595