Biochemical, Bioinformatics and Crystallographic Analyses of Type III Restriction-Modification Enzyme EcoP1I

A thesis submitted to a partial fulfilment of the requirement

for the degree of

DOCTOR OF PHILOSOPHY

IN BIOLOGY

by

Manasi Vilas Kulkarni

20103084

INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH, PUNE

2016

CERTIFICATE

Certified that the work incorporated in thesis titled '**Biochemical, Bioinformatics and Crystallographic Analyses of the Type III Restriction-Modification Enzyme EcoP1I**', submitted by **Manasi Vilas Kulkarni** was carried out by the candidate, under my supervision. The work presented here or any part of it has not been included in any other thesis submitted previously for the award of any degree or diploma from any other university or institution.

Date: 13/12/2016

Dr. Saikrishnan Kayarat

(Thesis Supervisor)

DECLARATION

I declare that this written submission represents my ideas in my own words and where others ideas have been included; I have adequately cited and referred to original sources. I also declare that I have adhered to all principles of academic honesty and integrity and I have not misinterpreted or fabricated or falsified any idea/data/fact/source in my submission. I understand that violation of the above can cause disciplinary action by the institute and evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Date: 13|12|2016

MKulkarni

Manasi Vilas Kulkarni

(Registration No. 20103084)

Acknowledgements

A PhD is a long haul with multitudes of hardships, troubleshoots and friendships that come along the way. Writing acknowledgements hence was one of the toughest tasks to do.

Firstly, I would like to express my sincere gratitude to my advisor Dr. Saikrishnan Kayarat for giving me a chance to work in his lab as his very first graduate student. Setting up the lab from scratch in itself was a colossal learning experience. I am grateful to him for critically evaluating my experiments and scientific thought behind each proposal and inculcating the essentials of "how to do science". I consider myself fortunate to have learnt all the basics from him "first hand" as he stood by me forbearingly all the time in my initial days of PhD. Unknowingly, he has made me a better individual, since each day of my PhD was spent in learning perseverance, determination, dedication and perfection from him.

I am thankful to my research advisory committee: Dr. C. G. Suresh (NCL, Pune), Dr Nagaraj Balsubramanian (IISER, Pune), and Dr. Deepak Barua (IISER, Pune), for their insightful comments and encouragement. I would like to extend my gratitude to Dr. Gayathri Pananghat for critically assessing my research during our lab-meetings and also otherwise. I also thank her for proofreading my thesis very cautiously and thoroughly.

I would like to thank Prof. L. S. Shashidhara (Professor and Chair, Biology, IISER, Pune) for his constant support right from my first day at IISER, Pune. Along with academic support, I was indebted to receive his affectionate help even in personal life such as arranging blood for my sister's open heart surgery in March 2011.

I thank Dr. Harinath Chakrapani and his graduate student Satish Malwal for letting me work in their lab for synthesizing Werner's reagent, an important crystallization condition additive, which ultimately helped me crystallize a 260 kDa huge protein. I sincerely thank Chaitanya Erady (BS-MS student, IISER, Pune) for carrying out absolutely clean ATPase assays which formed an essential component of my research.

Next, I extend my gratitude to the lab-members of SK lab and G3 lab for making my life fun-filled at the bench during difficult experiments. I would specifically thank Neha Nirwan, Ishtiyaq Ahmed Khan, Mahesh Kumar Chand, Shrikant Harne and Jyoti Baranwal for all the help and support. I would like to thank Mrinalini Virkar (Biomanager) and her team-members for uninterrupted supply of lab-essentials. I am also grateful to library officials Umeshareddy Kacherki and Tanuja Sapre for their prompt help for accessing journal papers.

I could not have survived without a few special people who joined my hands and stood by me practically and emotionally in all odds. I thank Niraja Bapat, Shubhankar Kulkarni, Shrikant Harne and Ketakee Ghate for providing me a cozy environment of long-lasting friendship. I would like to thank my batch-mates Vallari Shukla, Aparna Sherlekar, Devika Ranade, Rashmi Kulkarni and Bhavani Natarajan for their persistent encouragement and love.

Music was my true companion in dealing with burdens of PhD. Very special thanks to Abhijeet Bayani, Srikishna Sekhar and Trupti Thite for all the music jamming sessions that went on for hours together and helped me dissolve the stress of PhD in octaves of music. Beyond music, I share a long lasting, love-fun-filled friendship with this master-trio.

I would like to thank my family: my parents and my sister, Aditi for supporting me spiritually throughout. I could not have accomplished this thesis without their endless support as they constantly looked after my new-born baby while I wrote. I am also grateful to my parents in laws for their kind support and patience.

This acknowledgement cannot be completed without thanking my loving husband, Nachiket Khasnis who kept me motivated and never expected anything else from me over working at lab meticulously and honestly. His sacrifices are countless and words would dry up acknowledging all of them!

Manasi Vilas Kulkarni
December 2016

List of Abbreviations

°C	Degree Celsius
µL	Microliters
µm	Micrometers
µM	Micromolar
1D	1 Dimensional
3D	3 Dimensional
Å	Angstorm
AAA+	ATPases associated with diverse cellular Activities
ABC	ATP-binding cassette transporters
ABD	AdoMet binding domain
AdoMet	S adenosine methionine
ADP	Adenosine diphosphate
AHJR	Archeal Holliday Junction Resolvase
AMP	Adenosine monophosphate
AMP-PNP	Adenylyl-imidodiphosphate
ATP	Adenosine triphosphate
bp	Base Pair
BSA	Bovine serum albumin
cm	Centimeters
CTD	C-terminal domain
DCM	Dichloromethane
DNA	Deoxyribonucleic acid
DNAse	Deoxyribonuclease
dsDNA	Double stranded deoxyribonucleic acid
DTT	Dithiothreitol
E. coli	*Escherichia coli*
E:S	Enzyme:Site
EDTA	Ethylenediaminetetraacetic acid
EMSA	Electrophoeretic mobility shift assay
EOP	Efficiency of plating

g/L	grams per litre
GC	Guanine and cytosine
GTP	Guanosine triphosphate
Hep	Heparin
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
Hjc	Holliday Junction Resolvase
HsdM	Host specificity for DNA modification
HsdR	Host specificity for DNA restriction
HsdS	Host specificity for DNA specificity
HtH	Head to Head
HtT	Head to Tail
IDT	Integrated DNA technologies
iP	Inorganic phosphate
IPTG	Isopropyl β-D-1-thiogalactopyranoside
ISP	Type I Single Polypeptide
ITC	Isothermal calorimetry
K	Kelvin
kb	Kilobase
K_D	Dissociation constant
kDa	Kilodalton
L	Litre
LB	Lysogeny broth
LC/ESI-MS	Liquid Chromatography / Electrospray ionization mass spectrometry
Maldi-MS	Matrix Assisted Laser Desorption mass spectrometry
MDR	Modification Dependent Restriction
MES	2-(N-morpholine)-ethanesulfonic acid
mg	Milligrams
mL	Milliliters
mM	Millimolar
MPD	2-Methyl-2,4-pentanediol
Mrr	Methylated adenine recognition and restriction
MS	Mass spectroscopy
MSA	Multiple sequence alignment

MTases	Methyltransferase
MW	Molecular weight
MWCO	Molecular weight cut-off
NA	Nucleic acid
NDSB	Non-Detergents Sulfobetaines
NEB	New England Biolabs®
Ni-NTA	Nickel nitrilotriacetic acid
NJ	Neighbour joining
nm	Nanometers
nt	Nucleotide
NTD	N-terminal domain
NTP	Nucleotide triphosphate
OC	Open circular
OD	Optical density
PacBio	Pacific BioSciences
PAGE	Polyacrylamide gel electrophoresis
PC	Parent condition
PCR	Polymerase chain reaction
PEG	Polyethylene glycol
PSI-BLAST	Position-Specific Iterative Basic Local Alignment Search Tool
Py	Pyrimidine
REBASE	Restriction enzyme database
RecA	Recombinase A
RecB	Recombinase B
RF	Restriction free
RM	Restriction-Modification
RPM	Revolutions per minute
S	Second
SAXS	Small angle X-ray scattering
SC	Supercoiled
SDS	Sodium dodecyl sulphate
SEC	Size exclusion chromatography
SEC - MALS	Size exclusion chromatography + multi angle light scattering

Se-Met	Selenomethionine
SF	Sinefungine
SF 2	Superfamily 2
SMRT	Single Molecule Real Time
ssRNA	Single-stranded RNA
TBE	Tris borate EDTA
TIRF	Total internal reflection fluorescence
TLC	Translocation Looping & Collision
TRD	Target recognition domain
TtH	Tail to Head
TtT	Tail to Tail
UV-Vis	Ultra violet- visible
V_0	Initial velocity
Vsr	Very short patch repair

List of Tables

List of Figures

CHAPTER 1

Introduction

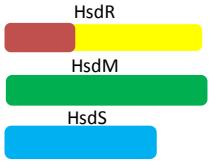
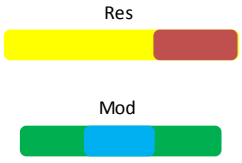
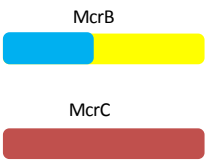
Chapter 1

Introduction

1. General Introduction

Protection against attack of bacteriophages is one of the main requirements for survival of bacteria (1–8). Bacteria have evolved a number of strategies to combat the invasion of foreign DNA, such as those from bacteriophages; the most prominent among them are Restriction Modification (RM) enzymes (9). These are molecular weapons used by bacteria to fight against viral attack. The history of RM enzymes dates back to early 1950s, when "host specificity" was described as a combination of two counteracting functions - endonucleolytic destruction (restriction) of foreign DNA and protection of self-DNA by base-specific modification (8, 10, 11). The RM enzymes are classified into four classes, Type I, II, III and IV, based on their subunit compositions, target site recognition, cleavage position and cofactor requirements (Table 1.1) (9, 12).

Table 1.1: Classification of NTP-dependent restriction enzymes

Feature	Type I	Type III	Type IV
Cofactors for Endonucleolytic Activity	Mg^{2+}, ATP	Mg^{2+}, ATP	Mg^{2+}, GTP
Cofactors for MTase Activity	AdoMet	AdoMet	-
DNA Target Site	Asymmetric, bipartite	Asymmetric, uninterrupted	Bipartite, RmC(N)40–80RmC
Subunits	HsdR (R) , HsdM (M) , HsdS (S)	Res, Mod	McrBL, McrBS, McrC
Domain Organization			
Position of DNA Cleavage	Far from target site, approximately in the middle of the two sites	Close to one of the target sites	Close to one of the target sites
Examples	EcoR124I	EcoP1I	McrBC

Type II RM enzymes are the simplest of all and consist of two separate enzymes for methylation and restriction, and do not depend on NTP hydrolysis for catalyzing the cleavage of unmodified DNA (13). These enzymes are used in DNA recombinant technology for controlled manipulation of DNA, such as cloning (14). Type I, III and IV are NTP-dependent, large multi-subunit assemblies, where the enzymatic activities of methylation and restriction are performed by different functional domains (15–18). The NTP dependent RM enzymes represent large molecular machines, which translocate on the double-stranded (ds) DNA utilizing energy from hydrolysis of either ATP or GTP (12). Type I and III RM enzymes catalyze the cleavage of unmodified DNA by utilizing free energy associated with ATP hydrolysis (19–22). Type IV systems are also called as Modification Dependent Restriction enzymes (MDR), since they restrict specifically methylated DNA in a GTP-dependent manner (23).

We focused on Type III RM enzymes to unravel the mechanistic basis of restriction and modification using the tools of biochemistry, X-ray crystallography and bioinformatics. Type III RM enzymes were discovered more than 40 years ago, and are now recognized as part of bacterial innate immune system. According to the latest information from Restriction Enzyme Database (REBASE) (24), more than 10,000 putative Type III RM enzymes (including only modification subunits) have been identified and target site sequences of more than 200 of them determined. Type III RM enzymes are hetero-oligomeric and multifunctional. They are composed of two subunits: Mod, which carries out site-specific methylation of DNA and Res, which catalyzes the cleavage of unmodified DNA, but only in complex with Mod (25). The target sequences of Type III RM enzymes are non-palindromic, asymmetric and uninterrupted (26–29). The cleavage requires a specific orientation of the target sites (30–32). The enzyme cleaves unmodified DNA 25-27 bp downstream of either one of the at least two inversely oriented target sites in an ATP-dependent fashion (20, 28, 29). The most studied prototypes of Type III RM enzymes are EcoP1I and EcoP15I. To understand the molecular basis of working of Type III RM enzymes, we focused our studies on EcoP1I.

This chapter describes historical milestones in the discovery of RM enzymes and introduces properties of Type III RM enzymes based on previous genetics, biochemical and biophysical studies.

2. Historical Background

The first-half of 20th century witnessed vital discoveries right from X-ray diffraction by Max von Laue in 1912 to the helical structure of DNA, the molecular basis of life by Watson and Crick in 1953 forming the underpinnings for identification of defense systems of bacteria, the RM enzymes. The most simplistic explanation of defense systems of bacteria was given by Nobel laureate Werner Arber's 10 years old daughter Silvia via "The tale of the king and his servants" as described in his biography (33).

"When I come to the laboratory of my father, I usually see some plates lying on the tables. These plates contain colonies of bacteria. These colonies remind me of a city with many inhabitants. In each bacterium there is a king. He is very long, but skinny. The king has many servants. These are thick and short, almost like balls. My father calls the king DNA, and the servants enzymes. The king is like a book, in which everything is noted on the work to be done by the servants. For us human beings these instructions of the king are a mystery.

My father has discovered a servant who serves as a pair of scissors. If a foreign king invades a bacterium, this servant can cut him in small fragments, but he does not do any harm to his own king.

Clever people use the servant with the scissors to find out the secrets of the kings. To do so, they collect many servants with scissors and put them onto a king, so that the king is cut into pieces. With the resulting little pieces it is much easier to investigate the secrets. For this reason my father received the Nobel Prize for the discovery of the servant with the scissors."

The RM enzymes were fondly referred to as "servant with the scissors" by Silvia. Following section describes the historical milestones in the discovery of these molecular scissors.

2.1 Host controlled variation and its molecular basis

Historical experiments by Luria & Human and Bertani & Weigle pioneered early discoveries of bacterial immune systems. It was observed that bacteriophage had a host specific capability of invading the bacteria wherein it could infect specific bacterial strain and failed to infect others. In 1952, Luria & Human (1) and Bertani & Weigle (2) independently experimented on λ phages and T even phages. They showed that it is a "host controlled variation" that caused difference in the bacteriophages. The host controlled variation in a bacteriophage was shown to be an effect of genotype of the host in which it grew. This change suppressed the infectious ability of the phage in other hosts. The authors demonstrated that the suppression of infection was transient and non-hereditary; meaning just a single passage on the previous host returned the phage to its original form. The Bertani and Weigle experiment is represented in Figure 1.1.

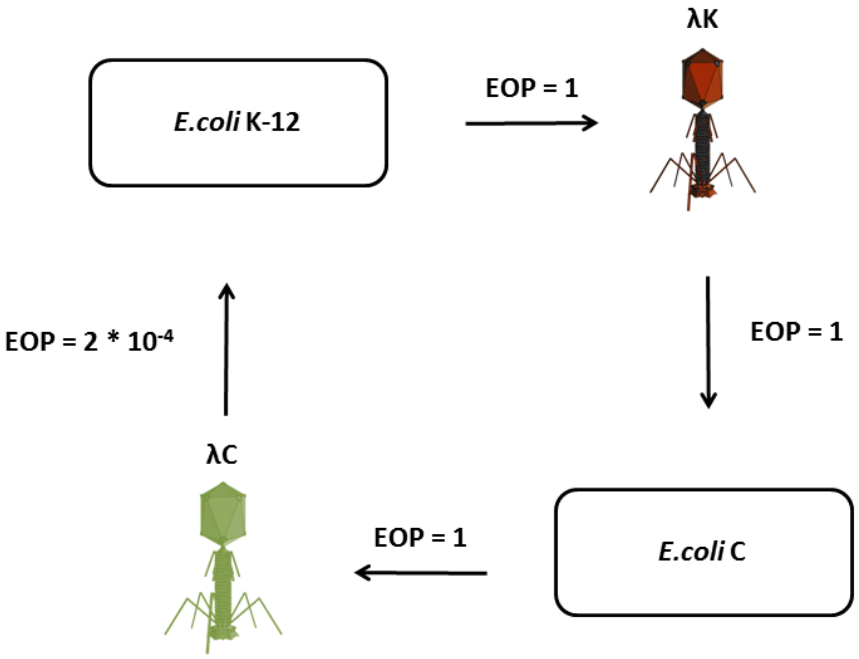


Figure 1.1: The Bertani and Weigle experiment of host controlled variation (2)**.** The λ phage propagated in the strain *E.coli* K12, referred to as λK, with an EOP (see text below) of 1. Subsequently the λ phage propagated in the strain *E.coli* C, referred to as λC, now faces a barrier to infect its original host *E.coli* K12 with an EOP of $2*10^{-4}$. However, phages overcoming this barrier return to their original form λK after a single passage in *E.coli* K12.

In 1960, in an attempt to render *E.coli* B and its radiation-resistant strain B/r sensitive to λ phage, Werner Arber encountered "host-controlled variation of phages"; a phenomenon described for *E.coli* and λ phages seven years earlier by Bertani & Weigle

and Luria & Human. To understand how restriction of phage growth and adaptation to new host worked, Arber and co-workers investigated this phenomenon in detail. On the basis of available knowledge on the structure of DNA (34) and of its semiconservative replication (35), they succeeded in showing that phage DNA carry host-strain specific modification and that unmodified DNA becomes degraded in the restricting host bacteria. The experiments were not just limited to understanding phenotypic implications of host-controlled variation, but went on to prove that the modification imparted by host, was associated with DNA (3, 4). They showed that the modification was present only in progeny phage particles which inherited one or both parental DNA strands of the infecting phage particle. Furthermore they showed that the unmodified phage DNA was degraded upon injecting in a restricting host (6, 7). Based on these findings, Arber and co-workers hypothesized that bacteria might contain two types of enzymes: a restriction enzyme that recognizes and cleaves the foreign bacteriophage DNA and a modification enzyme that recognizes and modifies the bacterial DNA to protect it from the DNA-degrading activity of its very own restriction enzyme (36). He predicted that the restriction enzyme and the modification enzyme act on the same DNA sequence, called a target/recognition sequence. In this way, the bacterial cell's own self-defense mechanism was identified, which destructively degrades invading phage DNA, and at the same time safeguards its own DNA from degradation by specific modification of the target sequence. The findings of Arber and co-workers led to the discovery of Type I RM enzymes EcoK and EcoB (3, 4, 6, 7, 36–38) and could provide explanation to early observations by Luria & Human and Bertani & Weigle. The success of λ phage infection in *E.coli* C in Bertani & Weigle's experiment (Figure 1.1) was due to absence of a RM enzyme in this strain as opposed to *E.coli* K12 which harbored EcoK RM enzyme.

EcoK and EcoB, the first RM enzymes to be discovered, were shown to be dependent on ATP, S adenosine methionine (AdoMet) and Mg^{2+} for enzymatic activities (36, 39) and were subsequently called as Type I RM enzymes. Soon in 1972, Horiuchi and Zinder showed that EcoB cleaves the DNA non-specifically significantly away from the target site (40). These initial discoveries of Type I RM enzymes were further corroborated by Hamilton Smith who also contributed to the discovery of Type II RM enzymes (41). Smith's findings were based on studies on restriction enzymes from *Haemophillus influenzeae*, where HindI was found to be a Type I RM enzyme while a second enzyme

from *H. influenzeae* (HindII) headed to appreciation of another class of enzymes, Type II RM enzymes, which cut the DNA at a specific location within the target site and do not depend on ATP for restricting the unmodified DNA (41). The Type II RM enzymes proved their importance in precise manipulation of pieces of DNA and opened new doors in the field of recombinant genetic technology.

2.2 Discovery of Type III RM enzymes

The RM enzymes identified so far were chromosomal. In parallel, the discovery of an extra-chromosomal independent RM enzyme stemmed from early experiments of Arber and Dussoix while trying to understand the molecular nature of restriction and modification. They experimented on λ phage and *E.coli* K12 (P1) lysogen. Lysogen is a bacterium harboring phage particles in dormant state (prophage P1 in this case) (42). The λ phage isolated from *E.coli* K12 was successfully restricted when plated on *E.coli* K12 (P1) lysogen with an efficiency of plating (EOP: number of plaques on the test host divided by the number of plaques on a permissive host) of 10^{-4}. However, when λ phage recovered from *E.coli* K12 (P1) lysogen was again plated on the same strain i.e. K12 (P1) lysogen, it was found to be infectious with an EOP of 1. One of the most striking observations was that the severity of infection was higher when phages recovered from K12 were plated on K12 (P1) lysogen than vice versa. This suggested that there were at least two independent RM systems in lysogenic K12 and only one in the non-lysogenic K12. Arber and co-workers showed that phage P1 had its specific modification which was distinct from previously identified chromosomal systems such as EcoK and EcoB (6–8).

Early experiments by Arber and co-workers established that the RM enzyme encoded on prophage P1 was extra-chromosomal as opposed to chromosomal RM enzymes such as EcoK (a Type I RM enzyme) (6). The discovery of this extra-chromosomal RM enzyme was followed by identification of a similar kind of RM enzyme in *E.coli* $15T^-$. In 1965, Stacey et al demonstrated by conjugation experiments, that *E.coli* $15T^-$ had a distinct, extra-chromosomal RM enzyme (8, 36). This RM enzyme was found to be present on the p15B plasmid of *E.coli* $15T^-$ (37). So far, two extra-chromosomal RM enzymes were identified: EcoP1I, which was carried by the P1 prophage, and EcoP15I, which was carried on a plasmid p15B. Both of these enzymes were shown to undergo genetic

recombination (11). They also had a competitive effect for stable inheritance (38). Interestingly it was observed that EcoP1I and EcoP15I depended on presence of ATP for successful restriction.

Later, in 1978, Piekarowicz and Kauc isolated a restriction enzyme from *Haemophilus influenzae*. This RM enzyme, HinfIII, had characteristics similar to EcoP1I and EcoP15I. It depended on the presence of ATP and Mg^{2+} for activity and was stimulated by AdoMet. These three enzymes lacked the huge ATP hydrolysis (1 ATP/bp) exhibited by Type I RM enzymes. Similar to Type II RM enzymes, these showed clear cut location of dsDNA break. Piekarowicz and Kauc proposed that EcoP1I, EcoP15I and HinfIII belonged to a separate, new class of RM enzymes: Type III RM enzymes (43).

3. Properties of Type III RM enzymes

Although a large number (>10,000) of putative Type III RM enzymes are known, only a handful of them have been characterized, including EcoP1I, EcoP15I, HinfIII, StyLI etc. (20, 27–29, 44). Among them, EcoP1I and EcoP15I have been studied most extensively for more than 40 years. Following section describes properties of Type III RM enzymes. This includes genetic make-up, subunit assembly, domain organization, sequence characteristics of various domains, cleavage characteristics and models proposed to explain working of these enzymes.

3.1 Genetic make-up

The information on genetic make-up of EcoP1I and EcoP15I RM stemmed from elegant experiments done by Mural et al and Iida et al. In 1979, Mural et al isolated and characterized cloned fragments of prophage P1 and showed that one of the recombinant plasmids expressed both modification and restriction activity of EcoP1I (45). Later in 1982, Iida et al identified the genetic make-up of EcoP1I and EcoP15I by using P1-P15 hybrid phages expressing EcoP15 target specificity. Iida and co-workers used combined information from transposon mutagenesis, restriction cleavage analysis (using BamHI and EcoRI), DNA heteroduplex analysis and *in vitro* transcription mapping to decipher the *mod-res* operon of EcoP1I and EcoP15I. These studies implied that the genetic information for both methylation and restriction was encoded by a single operon 5000 bp long (46). This genetic element encodes two genes: *mod* and *res*. Based on DNA heteroduplex analysis, Iida et al suggested that the *mod* gene of both

EcoP1I and EcoP15I was 2200 bp long. *res* of both EcoP1I and EcoP15I was 2800 bp long and was fairly identical for EcoP1I and EcoP15I. It was also suggested that each gene is transcribed from its own promoter (46). Later, in 1987, Huembelin et al reported the nucleotide sequence of EcoP1I operon and EcoP15I *mod* gene. With the availability of both EcoP1I and EcoP15I *mod* sequences, it was clear that the N and C termini of both the genes were highly homologous, but there existed a non-homologous region in the middle portion, making *mod* a "mosaic" of homologous and non-homologous regions (47).

Later, in 1979, the recognition sequences of EcoP1I and EcoP15I were determined. EcoP1I recognizes 5' AGACC 3', where the second adenine gets methylated by the modification activity (29). EcoP15I recognizes 5' CAGCAG 3', where again the second adenine gets methylated by the modification activity (28). Both the enzymes cleave 25-27 bp downstream of the target site leaving a 5' overhang in the cleaved product (11,12). Since EcoP1I and EcoP15I recognized two distinct stretches of nucleotides as their target site, it was suggested that presence of a non-homologous region in *mod* was due to different sequence specificities (47). The EcoP1I and EcoP15I operon are shown in Figure 1.2.

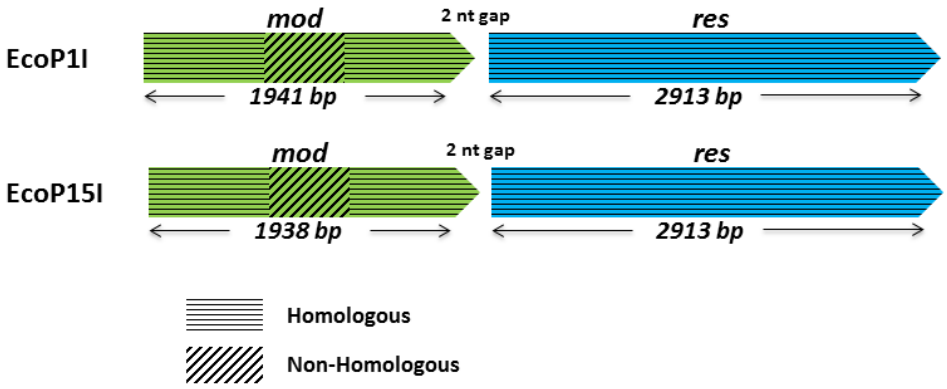


Figure 1.2: *mod-res* operon of EcoP1I and EcoP15I. The *mod* and *res* genes are shown as green and blue arrow-heads, respectively. The horizontal and oblique hash-marks denote regions of homology and non-homology, respectively.

3.2 Subunit assembly

In 1982, Hadi et al purified EcoP1I and EcoP15I from *E.coli* by a recombinant method where both *mod* and *res* were cloned on multicopy, overexpression plasmids (25). The authors reported that EcoP1I and EcoP15I contained two subunits of molecular weights 106 kDa and approximately 73-75 kDa respectively. Bacteria containing full-length operon of the restriction enzyme had r^+m^+ phenotype and produced both the subunits; however when a large portion of *res* was deleted, the bacterial phenotype was r^-m^+. The deletion derivatives expressed only the smaller subunit. The authors hence concluded that the smaller subunit was responsible for specific sequence recognition and methylation (25). In summary, Mod, the smaller subunit (~73-75 kDa) is transcribed from *mod* and is responsible for methylation of the target adenine. Res, the larger subunit (~106 kDa) is transcribed from *res* and is responsible for restriction of DNA when combined with Mod.

For a long time it was believed that Type III RM enzymes existed as a tetramer with a 1:1 stoichiometry of Mod:Res forming a hetero-tetrameric complex of Mod_2Res_2, and the active methyltransferase (MTase) as a dimer of two Mod subunits. These results were based on analytical ultracentrifugation and Size Exclusion Chromatography (SEC) (10, 48, 49). Wyszomirski et al revisited the subunit stoichiometry of EcoP15I in 2011. To dissect out the hetero-oligomeric nature of EcoP15I, they used three different techniques: SEC, second derivative UV spectroscopic analysis and analytical ultracentrifugation. Results obtained from all the three methods indicated that the multifunctional EcoP15I was a complex of 2 Mods and 1 Res forming a heterotrimer (Mod_2Res_1). The MTase was still found to be a homo-dimer (50). However, again in 2012, Gupta et al used analytical ultracentrifugation and dynamic light scattering experiments to show that EcoP15I was a heterotetramer. In the same report, a low resolution solution structure of EcoP15I using small angle X-ray scattering (SAXS) was reported. The authors suggested that EcoP15I had a crescent shape where two Mods lie in the middle and each of the Res subunits lay at the end (51).

In 2014, however, more detailed experiments towards understanding subunit stoichiometry of EcoP1I, EcoP15I and PstII were carried out. Researchers used two techniques which were not applied to these enzymes previously namely, Native Mass

Spectroscopy (Native MS) and Size Exclusion Chromatography combined with multi-angle light scattering (SEC-MALS). Results obtained from both the methods demonstrated that Type III RM enzymes contained two copies of Mod and a single copy of Res forming a heterotrimer Mod_2Res_1 (52). Finally a partial structure of EcoP15I was published in 2015, which beyond doubt showed that EcoP15I indeed has a Mod_2Res_1 stoichiometry (53). The structure revealed that two copies of Mod (Mod_A and Mod_B) perform asymmetric functions. Mod_A is involved in identifying the target sequence whereas Mod_B brings about methylation of DNA by transferring a methyl group from AdoMet to exocyclic N6 of adenine in the target site.

3.3 Target site of Type III RM enzymes

As demonstrated by Arber and Dussoix, the modification of phage particle acquired from the host was associated with DNA and RM enzymes were capable of distinguishing specific base pair sequences on the polynucleotide track (3, 4, 7). Brockes et al showed that EcoP1I methylated adenine residues using radiolabelled AdoMet. This was further corroborated by experiments done by Hattmann et al. using *in vitro* methylation of labeled oligonucleotides by EcoP1I Mod (54). The authors proposed that the product of methylation was a pentameric stretch of 5' AGACPy 3' (Py : C or T) where the second adenine got methylated. Further Baechi et al conducted experiments with SV40 DNA and EcoP1I to show that the target site for EcoP1I was 5' AGACC 3', once again showing that the second adenine got methylated (29). Hadi et al determined the target site for EcoP15I as 5' CAGCAG 3', where similar to EcoP1I, second adenine got methylated (28). Soon after this, in 1980, Piekarowiz et al identified the sequence recognized by HinfIII, another Type III RM system to be 5' CGAAT 3' (27). The first pair of isoschizomers was identified in 1982 with the discovery HineI which recognized the same target sequence as that of HinfIII (26).

With the advent of Single Molecule Real Time (SMRT) sequencing, developed by Pacific BioSciences (PacBio), target site sequences of a large number of MTases of Type III RM enzymes are now becoming available (55). Table 1.2 lists Type III MTases whose target sequences have been determined either biochemically or by SMRT sequencing. A typical target sequence of Type III RM enzymes is an uninterrupted, non-palindromic and asymmetric stretch of five to six base pairs. The asymmetry imparts a distinct polarity

to the site, where 5' and 3' termini of the target site are designated as tail and head respectively.

Table 1.2: Target sequences of MTases of Type III RM enzymes*

Name	Target Sequence (5'-3')	Name	Target Sequence (5'-3')	Name	Target Sequence (5'-3')	Name	Target Sequence (5'-3')	Name	Target Sequence (5'-3')
M.Aci16581I	AGGAG	M.EcoJA23PII	CACAG	M.KpnNIH30II	CGCATC	M.Sen641I	CAGAG	M.Sen18569III	CAGAG
M.Asa43001I	CGCAT	M.EcoJA65PII	CAGCAG	M.KpnNIH32I	CGCATC	M.Sen1080I	CAGAG	M.SenA46I	CAGAG
M.Asp2D2II	TCCAG	M.EcoPI	AGACC	M.KpnPC07II	CGCATC	M.Sen1175II	CAGAG	M.SenAZII	CAGAG
M.AspDUT2IV	CGAGG	M.EcoP15I	CAGCAG	M.KpnPC33II	CGCATC	M.Sen1387I	CAGAG	M.SenAbaI	CAGAG
M.Bce842I	CACAG	M.Esp3131I	GTTAAT	M.Mca25239I	GARAC	M.Sen1427I	CAGAG	M.SenAboI	CAGAG
M.Bce895I	CACAG	M.Fen1006I	GAVATC	M.Mha183IV	GTTAAT	M.Sen1655I	CAGAG	M.SenAnaII	CAGAG
M.Bce16656I	CACAG	M.Fla104114I	CCAAG	M.Mha807I	GTTAAT	M.Sen1676II	CAGAG	M.SenC1736I	CAGAG
M.Bce25416I	CACAG	M.FnoB1III	CGCC	M.MmyCI	TGAG	M.Sen1677II	CAGAG	M.SenC1765I	CAGAG
M.Bce22E1I	CACAG	M.FpsJII	CGCAG	M.Msp315II	CAGAAA	M.Sen1728II	CAGAG	M.SenC1808I	CAGAG
M.Bce7H2I	CACAG	M.FtnUIII	AGACC	M.MspCY2I	AGCGCC	M.Sen1735I	CAGAG	M.SenC1810I	CAGAG
M.BceK56I	CACAG	M.Gel16401II	GGACCG	M.Ngo3502I	ACACC	M.Sen1736I	CAGAG	M.SenJI	CAGAG
M.BceSI	CGAAG	M.Gli15749II	GACAT	M.NgoAX	CCACC	M.Sen1764II	CAGAG	M.Sen4481ORF6431	CAGAG
M.Bco7050I	GAWTC	M.GmeII	TCCAGG	M.NgoAXII	AGAAA	M.Sen1766I	CAGAG	M.SenSPBII	CAGAG
M.BcoSlacI	TAAATC	M.Gsp12I	GCCAT	M.NgoFA19II	CCACC	M.Sen1781I	CAGAG	M.SenTFI	CAGAG
M.BmaBMKI	CACAG	M.GspL21II	GACCA	M.NgoMX	CCACC	M.Sen1783I	CAGAG	M.SmoLIV	CGWAG
M.BmaBMZI	CACAG	M.Gst10I	GCCAT	M.Nme18I	ACACC	M.Sen1878I	CAGAG	M.SptAII	CAGAG
M.Bpe137I	AGCCGCC	M.Gth3570II	GCCAT	M.Nme77I	ACACC	M.Sen1880I	CAGAG	M.Ssc1360I	GGAG
M.Bpe564I	AGCCGCC	M.Hal24586II	GGAG	M.Nme579I	CCAGC	M.Sen1896I	CAGAG	M.Sty13348I	CAGAG
M.Bpe1920I	AGCCGCC	M.HbaIII	CGCAGC	M.Nme579II	ACACC	M.Sen1898I	CAGAG	M.StyLTI	CAGAG
M.Bps1651I	CACAG	M.HindVI	CGAAT	M.NmeMC58I	CGYAG	M.Sen1899I	CAGAG	M.SwoAII	GTCAGG
M.Bps7894I	CACAG	M.HinmI	CGAAT	M.Pac19842III	ACCAGG	M.Sen1903I	CAGAG	M.Tca7334III	GCCAT
M.BpsBEMI	CACAG	M.HinfIII	CGAAT	M.Pae41639I	GCCCAG	M.Sen1906I	CAGAG	M.TdeIV	CTAAT
M.BspRB39II	CACAG	M.Hma11271II	AGYATC	M.PbaVA2II	GGSAG	M.Sen1908I	CAGAG	M.TmeBIV	CGCC
M.Btr190II	RGTAAT	M.Hpy99XXI	GWCAY	M.Psp32OWI	GGAGC	M.Sen1910I	CAGAG	M.TpaRLIII	GMGAGC
M.Btr192II	ACATC	M.HpyAXI	GCAG	M.PstII	CTGATG	M.Sen1921II	CAGAG	M.Tph12270III	CAGAAA
M.CpeAV	VGACAT	M.HpyAXVII	TCAG	M.Rsp7II	AGACC	M.Sen1927I	CAGAG	M.TspX514II	GGCAS
M.CthIII	GTCAT	M.Kaq16071II	GGACT	M.RthAD2I	CGACC	M.Sen2050I	CAGAG	M1.VcoRE98II	CCCACC
M.CthLQ8II	CGACC	M.Kpn36II	CGCATC	M.SbaUI	CAGAG	M.Sen2064II	CAGAG	M2.VcoRE98II	GACATG
M.Dac11109I	GACGA	M.Kpn677I	CGCATC	M.Sbo268I	CAGAG	M.Sen2069I	CAGAG	M.Vna16374II	CCACCG
M.DfeII	GAGAAG	M.Kpn1097I	CGCATC	M.Sbo12419I	CAGAG	M.Sen3124I	CAGAG	M.Vpa2008II	GNAATC
M.DthLIII	CACC	M.Kpn1420I	CGCATC	M.Sen16II	CAGAG	M.Sen7378I	CAGAG	M.Xor86I	CCGAGG
M.Eba57I	GAGAG	M.Kpn32192I	CGCATC	M.Sen16III	CCGAG	M.Sen8391II	CAGAG	M.Xor256I	GGAGG
M.EclHC3I	ACGAAG	M.Kpn34618I	CGCATC	M.Sen158I	CAGAG	M.Sen8692I	CAGAG	M.Xor2286II	GGAGG
M.Eco3858I	GAGAC	M.Kpn38547I	CGCATC	M.Sen195I	CAGAG	M.Sen10384I	CAGAG	M.Xor7331III	GCCAGG
M.EcoCFTII	CACAG	M.KpnNIH1II	CGCATC	M.Sen255I	CAGAG	M.Sen10708I	CAGAG	M.Xor7341III	GCCAGG
M.EcoGVIII	ACCACC	M.KpnNIH10II	CGCATC	M.Sen318II	CAGAG	M.Sen13311I	CAGAG	M.XorPXI	CCGAGG
M.EcoJA17PII	CACAG	M.KpnNIH24I	CGCATC	M.Sen483II	CAGAG	M.Sen15791I	CAGAG	M.YinY228II	CCGAG

*Information retrieved from REBASE on 16-10-2016

3.4 Site orientation selectivity

Along with identification of target site of HinfIII, Piekarowiz et al also experimented with ColE1 DNA to characterize the methylation and cleavage pattern of HinfIII (27). They mapped 5 potential cleavage sites for HinfIII. The authors did not observe any cleavage of DNA containing single target site. It was then suggested that HinfIII needs multiple sites for successful DNA cleavage (20).

Further, it was observed that certain phages were refractory to cleavage by EcoP15I. Genomic DNA of both phage T7 and T3 carry recognition site for EcoP15I viz 5' CAGCAG

3'. However, T7 was found to be refractory and T3 was found to be susceptible to cleavage by EcoP15I. There was one prominent difference between the genetic materials of the two: The genomic DNA of phage T7 had target sites of EcoP15I arranged unidirectionally, whereas at least two were in inverted orientation in that of phage T3. These findings along with evidences from cleavage experiments done on M13 phage led the authors to propose that Type III RM enzymes require two inversely oriented target sites for cleaving the dsDNA. This was the strand bias model proposed for Type III RM enzymes (30). The model was also successful in answering the prevention of suicidal cleavage of self-genomic DNA by Type III RM enzymes. After methylation, only one strand of the DNA carries the methyl group. When such a DNA replicates, only one methyl group is inherited by the daughter DNA molecule, but the same site remains completely unmethylated in the other, making it a substrate for destructive endonuclease. This creates problems of suicidal cleavage by Type III RM enzymes. To overcome this situation, Type III RM enzymes were hypothesized to use strand bias, wherein, restriction occurs only when two unmodified target sites are in inverted orientation and directly repeated sites would not be suitable for endonucleolytic restriction but could yet be modified. This is also called as "site orientation selectivity", where a productive interaction between two protein complexes occurs only when they are in specific relative orientation (32).

As mentioned earlier, the Type III RM enzymes recognize a short, non-palindromic and asymmetric sequence on DNA. Given that there is a polarity associated with such asymmetric site, we can now have inverted repeats in two distinct arrangements: Head-to-Head (HtH) and Tail-to-Tail fashion (TtT) (Figure 1.3). DNA molecules bearing such inversely oriented sites are referred to as canonical substrates for cleavage by Type III RM enzymes.

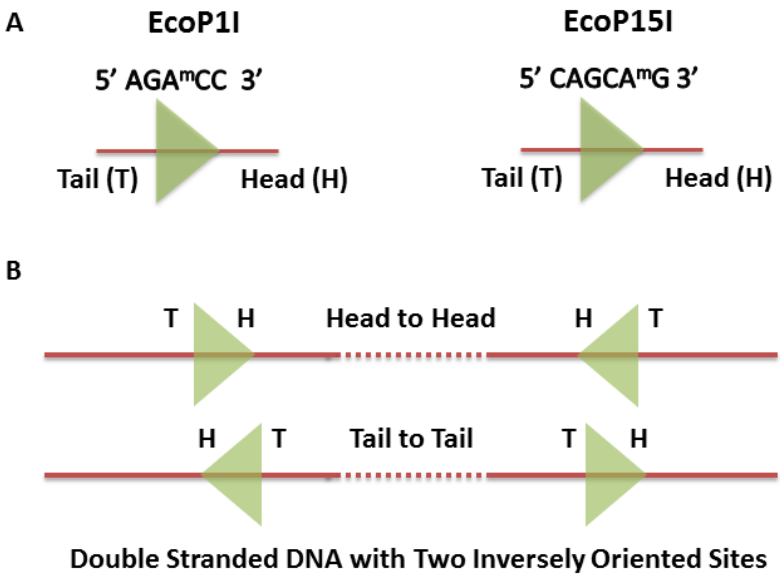


Figure 1.3: Site orientation selectivity. A] The asymmetric target sites for EcoP1I and EcoP15I are shown as triangles where the apex of the triangle denotes 3' end of target called as Head (H) and base of the triangle denotes 5' end of the target called as Tail (T). B] Indirectly repeated orientation of two target sites on a dsDNA.

3.5 Domain organization of Type III RM enzymes

Type III RM enzymes are composed of two subunits: Mod which is responsible for catalyzing transfer of a methyl group from AdoMet to exocyclic nitrogen of adenine; and Res which along with Mod brings about dsDNA break at a particular location (25-27 bp downstream of the recognition site). In 2007, Wagenfuehr et al subjected EcoP15I to limited proteolysis, mass spectrometry and insertional mutagenesis to characterize the structural domain within each subunit (15). In absence of specific DNA and ATP, Res exhibited two trypsin resistant domains of molecular weight 77-79 kDa (Res1) and 27-29 kDa (Res2). Such stable domains were not found upon proteolysis of Mod. To identify tightly folded parts of EcoP15I that resisted trypsin digestion, the Mod subunit and the tryptic fragements of Res namely, Res1 and Res2 were analyzed by a combination of liquid chromatography/ electrospray ionization mass spectrometry (LC/ESI-MS) and matrix assisted laser desorption/ionization mass spectrometry (Maldi-MS). The results from these experiments demonstrated that the Res subunit has two domains, a large N terminal domain that contains various helicase motifs and ATPase functional motifs, whereas the C terminal domain which was anticipated to be endonuclease domain. It was demonstrated that both these structural domains were connected by a linker region of 23 amino acids (15). Recently published three

dimensional structure of EcoP15I structure revealed presence of an additional accessory domain in EcoP15I Res, at the junction of RecA1 and RecA2, called the Pin domain (53).

Sequence characteristics of various functional domains within Type III RM enzymes are detailed in the following section.

3.5.1 Methyltransferase (Mod)

The Mod subunit of EcoP1I and EcoP15I is ~73-75 kDa (25, 48). Comparison of amino acid sequence of EcoP1I and EcoP15I Mod with other MTases brings up certain sequence motifs common to all MTases. MTases are broadly classified based on whether they catalyze a methyl group transfer to exocyclic amino groups of adenine (N6) or cytosine (N4) or to C5 atom of cytosine ring. Depending upon to which atom the methyl group transfer is catalyzed, these are either called N-MTases (transfer methyl group to exocyclic N of adenine or cytosine) or C-MTases (transfer methyl group to C5 of cytosine ring). Type III MTases were shown to transfer the methyl group from AdoMet to N6 of Adenine (29, 30) thus making them N-MTases. Comparison of N-MTases showed that there were two common motifs, but their linear arrangement at the primary sequence level and characteristics vary. According to the classification proposed by Timinskas et al, EcoP1I and EcoP15I belong to D21 class of N-MTases. Here D stands for a conserved aspartate (D) in motif II (DPPY). In the linear arrangement of motifs on the polypeptide chain of Type III Mod, motif II is located before the glycine rich motif I (FxGxG), hence this class is defined as D21 (56).

Later, Malone et al did a comprehensive structure guided comparative analysis of MTases and came up with a more elaborate classification (57). They compared the structures of M.HhaI a cytosine MTase and M.TaqI, an N6 adenine MTase. This structure guided comparison was supplemented with primary sequence analysis of N6 adenine, N4 cytosine and C5 cytosine MTases. Results of this analyses demonstrated that the MTases are composed of three functional domains: a catalytic domain, an AdoMet binding domain and a Target Recognition Domain (TRD). Each domain has characteristic motifs. The N-MTases are divided into three classes (α,ß and ϒ) based on the linear arrangement of 9 conserved motifs (57, 58).

Type III MTases belong to the ß class of MTases. Members of ß class have N terminal catalytic domains which have canonical motifs IV, V, VI, VII and VIII. The C terminal AdoMet binding domain has canonical motifs X, I, II and III. The TRD is interspaced between N terminal catalytic and C terminal AdoMet binding domain (Figure 1.4).

Figure 1.4: Domain organization of Mod of Type III RM enzymes. A] The Mod subunit is represented as combination of catalytic domain, AdoMet binding domain (pink blocks) and (TRD) (green block). B] Sequence alignment of EcoP1I, EcoP15I and BceSI depicting location of various motifs (red boxes). Roman numerals represent canonical motifs within ß MTases.

3.5.2 Restriction endonuclease (Res)

Res subunits of EcoP1I and EcoP15I are ~111 kDa (25). With the availability of amino acid sequences of Type III RM enzymes, Gorbalenya and Koonin reported that N-terminal part of Res subunit contained sequence motifs characteristic of Superfamily 2 (SF2) of DNA and RNA helicases (59). Helicases unwind nucleic acid duplexes by utilizing energy in the form of ATP hydrolysis (60). Based on the detection of signature motifs of SF2 helicases, it was postulated that Type III enzymes could also unwind the DNA duplexes (59). However, these enzymes did not show any strand separation activity (21). Type III enzymes thus are called pseudo-helicases which utilize energy obtained from ATP hydrolysis to communicate between recognition sites over long distances (30, 61) .

The N-terminal ATPase domain in Type III RM enzymes has a helicase core. This helicase core has two subdomains arranged in tandem. These two subdomains have a fold similar to Recombinase A (RecA), hence are also called RecA1 and RecA2 (60). Both RecA1 and RecA2 are involved in ATP dependent nucleic acid remodeling (62). The characteristic motifs of the helicase core can be classified based on their functions - 1) ATP binding and hydrolysis , 2) nucleic acid binding and 3) coordination between ATP and nucleic acid (60). Based on alignment of 39 amino acid sequences of Res subunits of Type III RM enzymes, McClelland and Szczelkun identified the core helicase motifs (63). Since structure of a Type III Res was unavailable then, a few motifs were declared as putative and their possible function predicted. Later, motifs Ib, Ic and IIa were identified by Gupta et al in 2015 based on the structure of EcoP15I (53). A schematic of location of various conserved canonical motifs based on information from predicted motifs from McClelland's and Szczelkun's study and partial structure of EcoP15I is depicted in Figure 1.5 (53, 63). As mentioned earlier an additional domain, called the Pin domain, is inserted between RecA1 and RecA2 of Res in EcoP15I (Figure 1.5A) (53).

Figure 1.5: Domain organization the Res subunit of Type III RM enzymes. A] The Res subunit of EcoP15I is represented as combination of RecA1, Pin, RecA2 (blue blocks) and endonuclease (grey block). B] Sequence alignment of EcoP1I, EcoP15I and BceSI depicting location of various motifs of SF2 helicases (red boxes) and AHJR nucleases (blue boxes). Roman numerals represent canonical motifs.

The endonuclease domain is at C-terminal of Res and has motifs characteristic of Archeal Holliday Junction Resolvase (AHJR) family of nucleases (64). The members of this family share a common fold with λ exonuclease and other types of endonucleases such as EcoRV (64–66). AHJR nucleases are characterized by a set of three conserved motifs I, II and III forming a catalytic triad (64). The nucleolytic activity is executed by a combination of these three motifs, where the conserved aspartate of motif II (PD) and, conserved glutamate, glutamine or aspartate in motif III (Q/E/DxK) coordinate with Mg^{2+} and one of the oxygens in the scissile phosphodiester bond in DNA. The conserved lysine in motif III contacts the phosphate of the DNA backbone (67). Motif I of the AHJR fold is characterized by a strongly conserved acidic residue (D/E). The acidic side chain in motif I has implications in stabilizing metal ion binding (68) or facilitating conformational transitions for coordination of the catalytic triad (69). Consensus sequence of motif I in Type III RM enzymes is still unidentified.

3.6 Mechanism of DNA cleavage by Type III RM enzymes

One of the striking features of Type III RM enzymes is the ability to communicate over large distances (> 1000 bp) utilizing only few tens of ATP molecules (22, 30, 61, 70). As described in Section 3.4, these enzymes maintain a sense of site orientation to bring about successful interaction between two enzyme complexes (30, 32, 61, 70, 71). For Type III RM enzymes it was suggested that HtH inverted repeat is favored over TtT (21, 30, 72). In parallel, it was observed that EcoP15I could also cleave TtT DNA substrates and that the cleavage efficiency was independent of the distance between the two sites (71). These findings were corroborated by van Aelst et al in 2010 (32). They experimented with EcoP1I, EcoP15I and PstII and showed that even TtT sites can be cleaved with equal efficiency. It was also demonstrated that the efficiency of cleavage depended on the lifetime of enzyme on the DNA (32). If the DNA were end-capped, the enzyme could stimulate cleavage of DNA with TtT sites with an equal efficiency. The authors suggested that cleavage of DNA containing both HtH and TtT sites is a common property of Type III RM enzymes. In the same study, communication between two protein complexes was shown to be necessary for cleavage. This was achieved by inserting a roadblock between two inverted target sites in both circular and linear DNA substrates. The effect of roadblock on the cleavage efficiency was little on circular substrates as compared to linear substrates. As the enzymes encountered a roadblock

on a circular substrate; they could still communicate with each other by other route on the circle. However, in a linear substrate, the enzymes could not communicate in an alternate way. This observation was consistent with a bidirectional, long range communication mode for Type III RM enzymes (32).

DNAs with a pair of inversely oriented sites, either HtH or TtT are canonical substrates for Type III RM enzymes (32). Cleavage of DNA with HtT or single target site was poor and hence cleavage of such substrates was referred to as secondary cleavage (73–76). Such substrates (HtT or single-site) will be called non-canonical substrates in this study. While a strict requirement for inversely oriented target sites was being observed for successful cleavage by Type III RM enzymes, there were observations regarding cleavage of above-mentioned non-canonical substrates. These were called "promiscuous cleavage events", and were believed to be an effect of enzyme concentration, cofactor requirement or the nature of monovalent cations in the reaction buffer. Potassium ions in the reaction buffer promoted the cleavage of single site plasmids; however sodium ions were shown to prevent promiscuity. Similarly, AdoMet was shown to inhibit such cleavage activity by EcoP1I (76). In another study, Raghavendra and Rao proposed requirement of a free 3' end for cleavage of DNA containing single or HtT sites. It was suggested that the translocating enzyme would interact with a free 3' end of the DNA. Such an interaction would trigger the reversal of direction of translocation on the DNA. The enzyme complex which is now translocating back towards the target site can interact with an enzyme which is already bound the site (75).

All of these observations were made on long stretches of DNA (>1000 bp). Interestingly, it was also observed that even smaller DNA fragments (50-70 bp) containing single target sites got efficiently cleaved by EcoP15I and EcoP1I (52, 71). These observations were hard to reconcile due to following findings:

1) The DNAseI footprint of EcoP15I is 36 nt (including the recognition site), where the enzyme covers 13 nt upstream and 17 nt downstream of the target site (71).

2) When a 50mer DNA duplex is complexed with either EcoP1I or EcoP15I, only one molecule of enzyme can bind one DNA molecule (52).

Thus, it is difficult to imagine two enzyme complexes interacting with each other on a small piece of DNA to bring about dsDNA cleavage. Further investigation is required to get more insights into such a cleavage pattern.

To summarize, Type III RM enzymes communicate between two inversely oriented sites over long distances (0-1000bp), wherein two protein complexes would physically interact with each other to elicit cleavage of the dsDNA 25-27 bp downstream of the target sites (21, 30, 32, 61, 71, 72). The entire activity requires hydrolysis of ATP; however, the amount of ATP hydrolyzed is 1000 fold less than other closely related NTP dependent enzymes such as Type I and Type ISP RM enzymes (19, 21, 22, 77). To account for site orientation selectivity along with extremely low ATP hydrolysis activity associated with Type III RM enzymes, different modes of communication are proposed (18, 32, 70, 75, 78–81).

3.6.1 Translocation, Looping and Collision (TLC) model

EcoK, EcoA and EcoB were among the first RM enzymes (Type I RM enzymes) to be identified. Type III RM enzymes share a homologous helicase core in the Res subunit with HsdR subunit of Type I RM enzymes. The translocation, looping and collision (TLC) proposed for Type III RM enzymes (Figure 1.6) borrowed its inspiration from that proposed for operation of Type I RM enzymes (82). In this model, after loading onto the HtH oriented recognition sites, the enzyme complexes start pulling the DNA towards them. As the length of pulled DNA goes on increasing, the distance between enzymes bound to the two sites goes on shortening. This eventually brings the two DNA bound enzyme complexes close to each other facilitating physical interaction between them. The physical interaction between two enzymes triggers the endonuclease activity of the enzyme and a dsDNA break is brought about. In this process, each base pair translocated on the DNA requires energy obtained from ATP hydrolysis yielding a coupling ratio of 1 ATP hydrolyzed/ base pair (12, 83). However Type III RM enzymes are much more energy efficient than the closely related Type I RM enzymes. They require as low as 1000 fold less ATP molecules to bring in the same biological effect, that is, to communicate between two distantly located sites through DNA translocation, and then cleaving the dsDNA. The TLC model fails to account for such a low ATP coupling ratio.

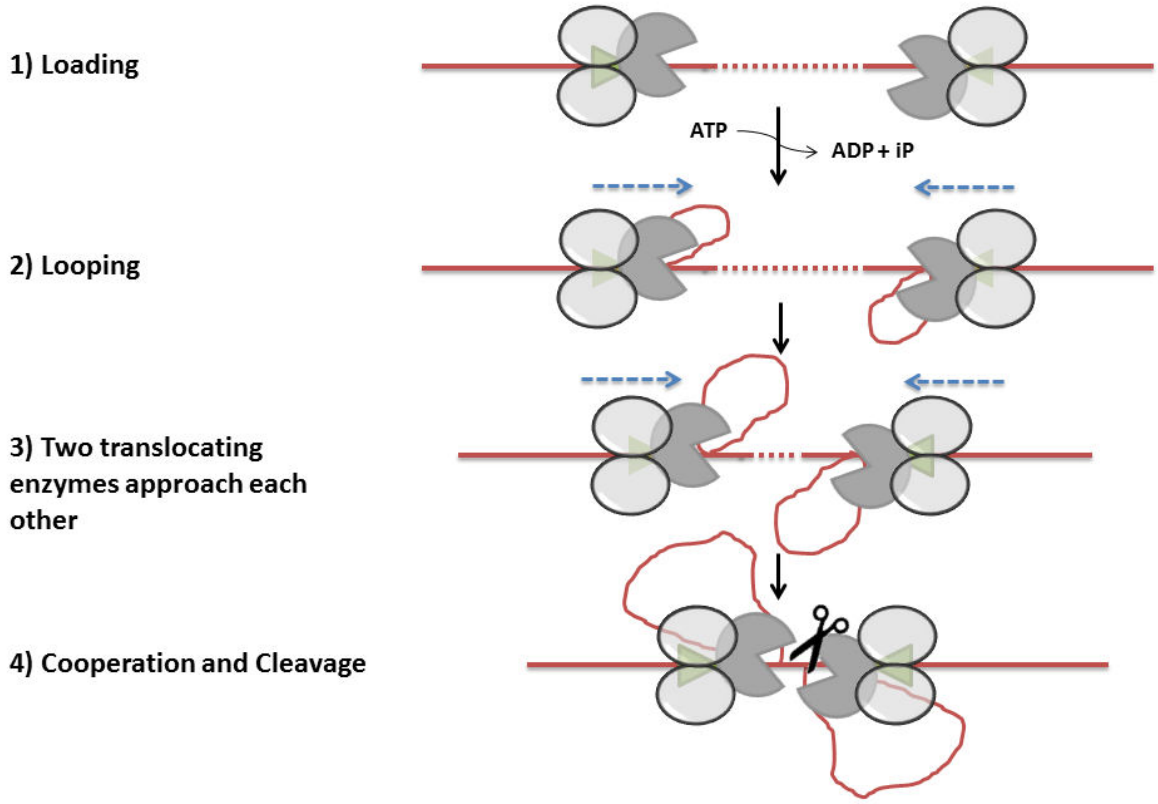


Figure 1.6: Translocation, Looping and Collision (TLC) model. DNA is shown as red line. The target site is shown as triangles on the DNA. The grey closed ovals represent Mod_1 and Mod_2 while the oval with a mouth represents Res. In the first step, the enzymes load onto the target site. Subsequent to successful loading, the enzymes pull the DNA utilizing ATP. The process of looping shortens the length between two translocating enzymes bringing them closer. Collision of two translocating enzymes triggers endonucleolytic cleavage of dsDNA.

3.6.2 End reversal model

As mentioned in Section 3.6, apart from cleavage of canonical substrates (HtH or TtT), cleavage of DNA containing HtT sites and single site was also observed (75, 76). To account for how single site linear substrates could be cleaved by Type III RM enzymes, the importance of a free DNA end was proposed (75). In this model, the enzyme complex loads onto the recognition site. Upon hydrolysis of ATP, the enzyme vacates the recognition site and starts translocating towards the 3' end. As the enzyme encounters a DNA end, it changes its direction of translocation. The subunit stoichiometry of Type III RM enzymes was believed to be Mod_2Res_2 at the time of this hypothesis. Hence researchers suggested that the reversal of translocation could occur by one of the two means: 1) The motor of second Res subunit is activated; 2) The enzyme makes a 180°

turn upon interaction with DNA end. The back-traversing enzyme now can approach another site bound enzyme. Physical interaction between the two enzymes activates the endonuclease to cleave the DNA (Figure 1.7). In this way, the model was able to explain cleavage of DNA substrates containing HtH, TtT and also HtT substrates. This would mean that direct and inverse repeats and single site substrates could be cleaved with same efficiency. Literature, however, suggests that the efficiency of cleavage to be higher for inverse repeats as compared to other orientations. Also, this model fails to account for cleavage of circular DNA having a single target site. On circular DNA with single target site, the enzyme cannot find an end to reverse the direction of translocation, making it simply impossible to juxtapose two enzyme complexes. Hence, this model also is unable to provide a concrete explanation for all the cleavage characteristics of Type III RM enzymes.

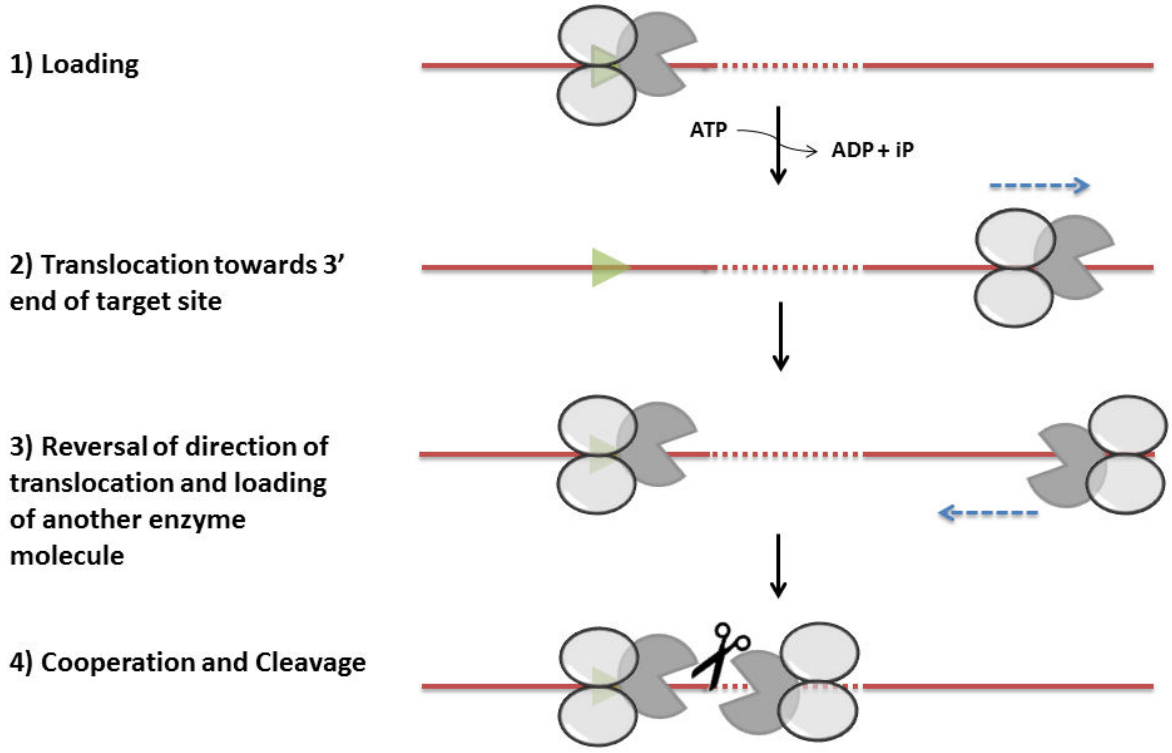


Figure 1.7: End reversal model. Enzyme loads on the DNA at the target site. Hydrolysis of ATP triggers translocation of the enzyme towards 3' end of the target site; thereby enzyme leaves the target site. In the meantime, this vacant target site gets occupied by another enzyme. The enzyme travelling towards 3' end reverses its direction of translocation and approaches the target site bound enzyme. Collision of two enzymes results in endonucleolytic cleavage of dsDNA.

3.6.3 Transient looping and collision model

In 2007, fast scan atomic force microscopy was done on EcoP15I with its substrate DNA to visualize in real time the DNA processing by Type III RM enzymes. EcoP15I was shown to form loops while still bound to the target site. Additionally, the enzyme contacted the non-specific stretch on the DNA beyond the loop. The DNA loops were completely absent on DNA substrates having no target sites. Based on the observations of atomic force microscopy, another model was proposed which tried to account for communication by Type III RM enzymes over large distances, yet hydrolyzing very less amount of ATP (Figure 1.8). According to this model, the enzyme first loads on the recognition site. ATP hydrolysis then brings about a conformational change in the enzyme, such that the ATPase domain of Res subunit can now transiently hold or release the DNA segment leading to formation of DNA loops in 3D space. Such a 3D DNA looping considerably reduces the distance between two distantly located site bound enzyme complexes. This looping was proposed to be diffusive, requiring no ATP hydrolysis. When the two enzyme complexes are sufficiently close to each other, a limited ATP driven active translocation brings them together to achieve physical interaction. Although the model could explain low ATP requirements of Type III RM enzymes, it could not account for cleavage of DNA containing very closely spaced target sites and single site. Enzyme dimers were not observed on single site substrates under the experimental conditions (80, 81).

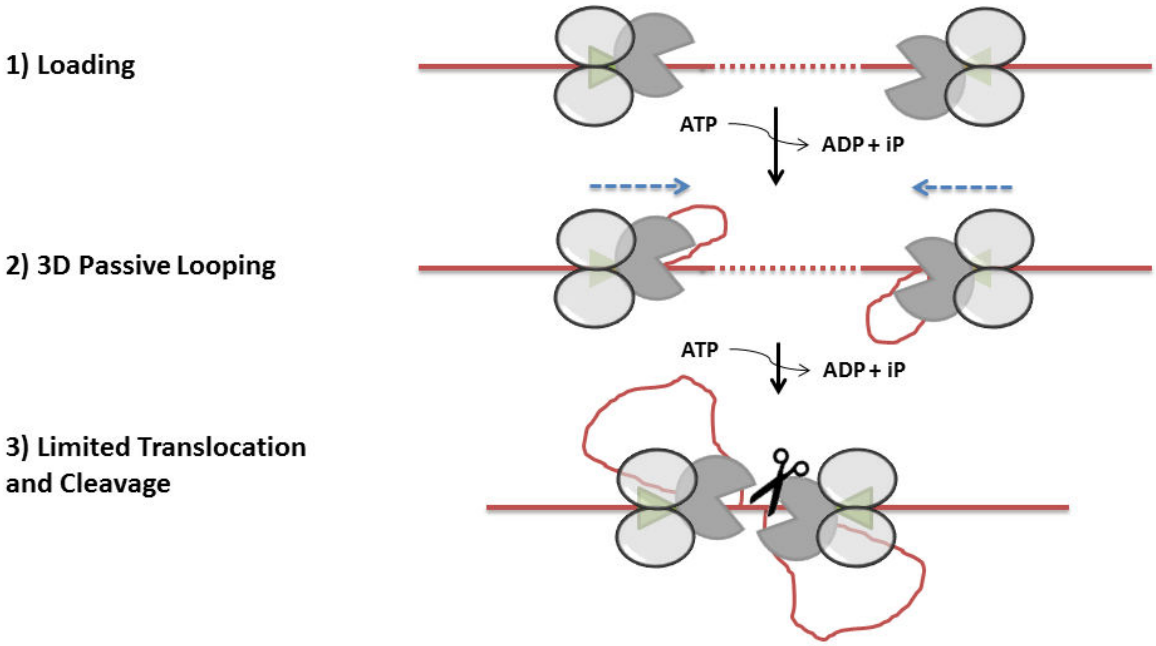


Figure 1.8: Transient looping and collision model. Upon loading on two inversely oriented target sites, ATP hydrolysis brings about a conformational change in the enzyme to facilitate passive 3D looping of the DNA segment between two enzymes. As the enzymes approach each other while still bound to site, a limited amount of ATP hydrolysis allows translocation of the enzymes resulting in collision and nucleolytic cleavage of dsDNA.

3.6.4 *1D bidirectional diffusion model*

The atomic force microscopy measurements pose technical challenge to the interpretation of data, since immobilization of DNA on mica can introduce bias in conformational flexibility of the enzyme. In 2010, Ramanathan et al used magnetic tweezers assays to observe the DNA processing behavior of EcoP1I and EcoP15I. In this method, one end of the DNA molecule is attached to the glass flow cell, while the other end is equipped with a magnetic bead and is held upright in the magnetic field. The DNA end to end distance is monitored in real time in presence or absence of a DNA interacting protein under study. If the DNA were looped by EcoP1I and EcoP15I, it would result in shortening of the DNA length. With EcoP1I and EcoP15I, no such change in DNA length was observed, however, the magnetic bead was lost from the field indicating the cleavage of DNA by EcoP1I subsequently. This led authors to propose that EcoP1I cleaved DNA without looping it.

Given that both EcoP1I and EcoP15I did not loop DNA and bring about cleavage by utilizing much less ATP, 1D diffusion was hypothesized to be the model for action of

Type III enzyme. According to this model, the enzymes load on the target site. The polarity of target site sets the polarity of the loaded enzyme. After the enzyme loads on the site, it undergoes an ATP dependent conformational change which enables it to slide past the site either upstream or downstream. The diffusing enzyme on encountering another target site-bound enzyme, the physical interaction between the two leads to activation of the endonuclease to bring about dsDNA cleavage. The cleavage step may be accompanied by another round of ATP hydrolysis. The model invokes requirement of ATP hydrolysis for merely loading on the site and not for inter-site communication. As bidirectionality is suggested for diffusing enzyme, the model also answers for cleavage of TtT substrates (32, 70).

Later in 2013, Schwarz et al used magnetic tweezers combined with total internal reflection fluorescence (TIRF) to monitor the mode of communication of EcoP15I. The authors reported that EcoP15I diffused on the DNA both upstream and downstream of the recognition site with a diffusion constant of ~ 0.92 $\mu m^2 s^{-1}$. It was also shown that ATP abolished the non-specific binding of EcoP15I and triggered the release of enzyme from its target site. Analogs of ATP like AMP-PNP, ADP-Vanadate and ADP did not abolish non-specific binding events, confirming that ATP acts as a conformational switch to push the enzyme from loading mode to a much distinct sliding mode (61).

The results of magnetic tweezers combined with TIRF were further corroborated by millisecond time resolution stopped flow assays (22). It was demonstrated that EcoP1I and EcoP15I rapidly hydrolyzed ~10 ATPs in ~1 second upon identification of target site on short DNA duplexes. This burst phase was coupled to protein conformational change measured by changes in tryptophan fluorescence. The conformational switch was followed by a slower burst phase where ~20 ATPs were consumed. Although 1D bidirectional diffusion model satisfactorily explains the cleavage of inverted sites which are very close to each other, it fails to answer for cleavage of single site linear and circular substrates.

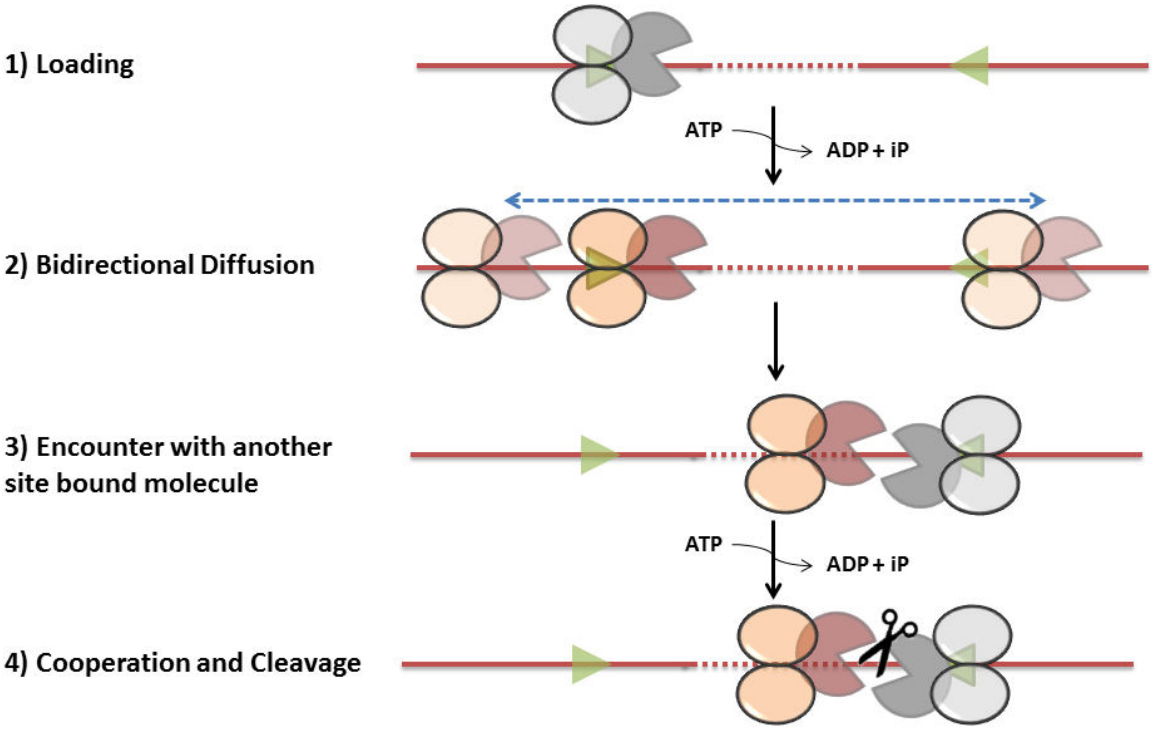


Figure 1.9: Bidirectional diffusion model. The enzyme loads on the target site. ATP hydrolysis brings about a conformational change in enzyme from loading (grey enzyme assembly) to sliding mode (colored enzyme assembly). The sliding enzyme can move past the target site bidirectionally. When such a diffusing enzyme encounters another target site bound enzyme, nucleolytic cleavage of dsDNA takes place possibly requiring another round of ATP hydrolysis.

4. EcoP1I: a prototype of Type III RM enzymes

To understand mechanistic basis of restriction and modification by Type III RM enzymes we chose to study EcoP1I, a prototypical Type III RM enzyme. As detailed in Section 2, EcoP1I was the first Type III RM enzyme to be identified (5, 10, 54, 84). EcoP1I is an extra-chromosomal system, where the contiguous gene segment of *res-mod* operon is carried on prophage P1. The operon of EcoP1I is 4856 bp long. *mod* is 1941 bp long, while *res* is 2913 bp long. There is a 2 bp gap between stop codon of *mod* and start codon of *res* (Figure 1.1) (25, 47). Both *res* and *mod* genes are transcribed from separate promoters. The promoters are located 70 bp and 140 bp upstream of the translation start codon of *mod* and *res*, respectively. There is an additional pair of promoters 500 bp upstream of the start codon of *mod*. The *res* gene is transcribed from within open reading frame of *mod* (47).

There are two subunits, viz, Mod (646 amino acids, MW: 73.48 kDa) and Res (970 amino acids, MW: 111.45 kDa), which are products of *mod* and *res*, respectively (25, 47). As described in Section 3.2, the enzyme assembly consists of two copies of Mod and a single copy of Res forming a hetero-trimer of Mod_2Res_1 (MW: 258.41 kDa)(52, 53). The target site for the enzyme is 5' AGACC 3'. The second adenine within the target site gets methylated by Mod, where it transfers the methyl group from AdoMet to exocyclic N6 of adenine in target site (29). The target site for EcoP1I is found in the genome of variety of viruses including SV40 (4 sites), λ virus (49 sites), ϕx 174 (7 sites) and adenovirus (78 sites)(24). Upon identification of a pair of inversely oriented unmodified target sites, the enzyme cleaves the DNA 25-27 bp downstream of the target sites by leaving a 5' overhang of 2 base pairs (29). ATP and magnesium ions are compulsory for endonucleolytic cleavage (21, 29).

As mentioned earlier, a piece of DNA containing inverse orientation of target sites (HtH or TtT) is referred to as canonical substrate for EcoP1I. All other substrates, such as those containing a pair of directly repeated target sites (HtT) or single site are non-canonical substrates (52, 73). As detailed in Section 3.6, cleavage models invoked to satisfy site orientation selectivity of Type III RM enzymes and low ATP consumption fail to explain cleavage of such substrates.

4. Summary

Type III RM enzymes were discovered more than 40 years ago. As more and more biochemical information was obtained, these enzymes were classified under a separate category, the Type III RM enzymes. This separate class of enzymes was characterized by two properties: 1) The enzymes cleaved the DNA at a very specific location (like a Type II RM enzyme) 2) The enzymes utilized quite low amounts of ATP to communicate between distant sites and to cleave the DNA. EcoP1I and EcoP15I are the most extensively studied prototypes of Type III RM enzymes. The enzymes are coded on an operon, containing 2 structural genes arranged in tandem. The methyltransferase (Mod,~73 kDa) is transcribed from *mod* (1900 bp), and the restriction endonuclease (Res, ~110 kDa) is transcribed from *res* (2900 bp). The active form of methyltransferase is a dimer of Mod, however, the restriction endonuclease is active only when associated with Mod forming a Mod_2Res_1 heterotrimer. Each subunit has different functional modules in the form of various domains. The Mod subunit belongs to ß class of MTases and has an N terminal catalytic and C terminal AdoMet binding domain. The TRD is inserted between catalytic and AdoMet binding domains. The Res subunit has an N terminal ATPase domain which belongs to SF2 helicases. The C terminal endonuclease domain belongs to the AHJR family of nucleases. The enzyme recognizes its target via TRD of one of the Mods. The other Mod is responsible for catalyzing transfer of a methyl group from AdoMet to the adenine of target site. For successful cleavage of non-methylated DNA, the enzyme requires two target sites in an inverted orientation which could be thousands of base pairs away, however, cleavage of substrates having directly repeated sites, very closely spaced sites or single site has also been observed. Interestingly, these enzymes communicate between a pair of sites utilizing very less amount of ATP, but still maintain the sense of site orientation. Different models have been proposed to explain the mechanism of action of Type III RM enzymes. All of these models, however, are only able to explain a few but not all characteristics of these complex systems. To summarize, in spite of availability of large number of experimental evidences, we still lack a complete picture of mechanism of action of these giant macromolecular machines.

5. Scope of Thesis

Type III RM enzymes, by virtue of their biochemical characteristics, represent model enzymes to study the working of large macromolecular machines involved in nucleic acid transactions. Several biochemical and biophysical studies have revealed interesting features about the mechanisms of action of Type III RM enzymes. However, structural information on the complexes has been limited due to the large size and nature of the enzyme complex. An ensemble of biochemical, bioinformatics and structural studies will provide a detailed view of the underlying mechanism.

Chapter 2: To determine the structure of Type III RM enzymes, we chose EcoP1I as the prototype. We started by cloning the gene under high expression vectors to get overexpression of the protein. This was followed by purification of the protein at large scale employing a new purification protocol to obtain sufficient amount of protein suitable primarily to set up crystallization trials. The enzymatic activities of EcoP1I purified by this new method were checked using DNA binding, ATP hydrolysis and DNA cleavage assays.

Chapter 3: To see the affinity of our highly purified and homogeneous protein sample towards DNA, we conducted DNA binding assays with various lengths of DNA substrates. During these trials, we characterized the binding and cleavage properties of EcoP1I in the presence or absence of various cofactors. Interestingly, we came across the ability of protein to cleave the DNA substrates containing single target site. We investigated in detail, the characteristics of such a cleavage. With the use of tools of biochemistry, we have proposed a model which explains the mechanism of such type of cleavage.

Chapter 4: To understand Type III RM enzymes in detail at the primary sequence level, we took up an extensive sequence analyses of these enzymes. Through these analyses along with identification of canonical motifs, we have been able to identify certain unique features of Type III RM enzymes at the primary sequence level. Finally, the newly identified sequence features were mapped on the structure to find out their functional relevance. This led us to identify new and unique motifs in Type III RM enzymes that are required for interaction with different cofactors.

Chapter 5: In parallel to characterization of EcoP1I purified in lab (Chapter 2), we crystallized EcoP1I with a suitable short DNA substrate mimic. After screening a large number of crystals, one of them diffracted to 4.4 Å. The diffraction data is anisotropic. However, we have successfully done molecular replacement by using partial structure of EcoP15I which was solved recently (53).

References

1. Luria,S.E. and Human,M.L. (1952) A nonheredetary, host-induced variation of bacterial viruses. *J. Bacteriol.*, **64**, 557–569.

2. Bertani,G. and Weigle,J.J. (1953) Host controlled variation in bacterial viruses. *J. Bacteriol.*, **65**, 113–121.

3. Arber,W. and Dussoix,D. (1962) Host specificity of DNA produced by Escherichia coli. I. Host controlled modification of bacteriophage lambda. *J. Mol. Biol.*, **5**, 18–36.

4. Dussoix,D. and Arber,W. (1962) Host specificity of DNA produced by Escherichia coli. II. Control over acceptance of DNA from infecting phage lambda. *J. Mol. Biol.*, **5**, 37–49.

5. Hattman,S. (1964) The control of host-induced modification by phage P1. *Virology*, **23**, 270–271.

6. Arber,W. and Morse,M.L. (1965) Host specificity of DNA produced by Escherichia coli. VI. Effects on bacterial conjugation. *Genetics*, **51**, 137–148.

7. Dussoix,D. and Arber,W. (1965) Host specificity of DNA produced by Escherichia coli IV . Host specificity of infectious DNA from bacteriophage lambda. *J. Mol. Biol.*, **11**, 238–246.

8. Stacey,K.A. (1965) Intracellular modification of nucleic acids. *Br. Med. Bull.*, **21**, 211–216.

9. Boyer,H. (1971) DNA restriction and modification mechanisms in bacteria. *Annu. Rev. Microbiol.*

10. Brockes,J.P. (1973) The deoxyribonucleic acid-modification enzyme of bacteriophage P1. *Biochem. J.*, **133**, 629–633.

11. Arber,W. and Linn,S. (1969) DNA modification and restriction. *Annu. Rev. Biochem.*, **38**, 467–500.

12. Dryden,D.T.F., Murray,N.E. and Rao,D.N. (2001) Nucleoside triphosphate-dependent restriction enzymes. *Nucleic Acids Res.*, **29**, 3728–3741.

13. Wilson,G.G. and Murray,N.E. (1991) Restriction and modification systems. *Annu. Rev. Genet.*, **25**, 585–627.

14. Nathans,D. and Smith,H.O. (1975) Restriction endonucleases in the analysis and restructuring of dna molecules. *Annu. Rev. Biochem.*, **44**, 273–293.

15. Wagenführ,K., Pieper,S., Mackeldanz,P., Linscheid,M., Krüger,D.H. and Reuter,M. (2007) Structural domains in the type III restriction endonuclease EcoP15I: characterization by limited proteolysis, mass spectrometry and insertional mutagenesis. *J. Mol. Biol.*, **366**, 93–102.

16. Rao,D.N., Saha,S. and Krishnamurthy,V. (2000) ATP-dependent restriction enzymes. *Prog. Nucleic Acid Res. Mol. Biol.*, **64**, 1–63.

17. Loenen,W.A.M., Dryden,D.T.F., Raleigh,E.A., Wilson,G.G. and Murray,N.E. (2014) Highlights of the DNA cutters : a short history of the restriction enzymes. *Nucleic Acids Res.*, **42**, 3–19.

18. Rao,D.N., Dryden,D.T.F. and Bheemanaik,S. (2013) Type III restriction-modification enzymes : a historical perspective. *Nucleic Acids Res.*, 10.1093/nar/gkt616.

19. Reiser,J. and Yuan,R. (1976) Purification and properties of the P15 specific restriction endonuclease from Escherichia coli. *J. Biol. Chem.*, **252**, 451–456.

20. Piekarowicz,A. and Brzeziński,R. (1980) Cleavage and methylation of DNA by the restriction endonuclease HinfIII isolated from Haemophilus influenzae Rf. *J. Mol. Biol.*, **144**, 415–429.

21. Saha,S. and Rao,D.N. (1995) ATP hydrolysis is required for DNA cleavage by EcoPI restriction enzyme. *J. Mol. Biol.*, **247**, 559–567.

22. Toth, J., Bollins,J. and Szczelkun,M.D. (2015) Re-evaluating the kinetics of ATP hydrolysis during initiation of DNA sliding by Type III restriction enzymes. *Nucleic Acids Res.*, **43**, 10870–10881.

23. Loenen,W.A.M. and Raleigh,E.A. (2014) The other face of restriction: modification-dependent enzymes. *Nucleic Acids Res.*, **42**, 56–69.

24. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2015) REBASE--a database for DNA

restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, **43**, D298-9.

25. Hadi,S.M., Bachi,B., Iida,S. and Bickle,T.A. (1983) DNA restriction-modification enzymes of phage P1 and plasmid p15B. Subunit functions and structural homologies. *J. Mol. Biol.*, **165**, 19–34.

26. Piekarowicz,A. (1982) HineI is an isoschizomer of HinfIII restriction endonuclease. *J. Mol. Biol.*, **157**, 373–381.

27. Piekarowicz,A., Bickle,T.A., Shepherd,J.C.W. and Ineichen,K. (1981) The DNA sequence recognised by the HinfIII restriction endonuclease. *J. Mol. Biol.*, **146**, 167–172.

28. Hadi,S.M., Baechi,B., Shepherd,J.C.W., Yuan,R., Ineichen,K. and Bickle,T.A. (1979) DNA Recognition and cleavage by the EcoPl5 restriction endonuclease. *J. Mol. Biol.*, **134**, 655–666.

29. Baechi,B., Reiser,J. and Pirrota,V. (1979) Methylation and cleavage sequences of the EcoPl enzyme. *J. Mol. Biol.*, **128**, 143–163.

30. Meisel,A., Bicklel,T.A., Kruger,D.H. and Schroeder,C. (1992) Type III restriction enzymes need two inversely oriented recognition sites for DNA cleavage. *Nature*, **355**, 467–469.

31. Kruger,D.H., Kupper,D., Meisel,A., Tierlich,M., Reuter,M. and Schroeder,C. (1995) Restriction endonucleases functionally interacting with two DNA sites. *Gene*, **157**, 165.

32. Aelst,K. Van, Tóth,J., Ramanathan,S.P., Schwarz,F.W., Seidel,R. and Szczelkun,M.D. (2010) Type III restriction enzymes cleave DNA by long-range interaction between sites in both head-to-head and tail-to-tail inverted repeat. *PNAS*, **107**, 9123–9128.

33. Nobelprize.org. Nobel Media AB 2014. Web. 8 Nov 2016. 'Werner Arber - Biographical'.<*http://www.nobelprize.org/nobel_prizes/medicine/laureates/1978/arber-bio.html*>,.

34. Watson,J.D. and Crick,F.H.C. (1953) Molecular structure of nucleic acids: A structure

of deoxyribose nucleic acid. *Nature*, **171**, 737–738.

35. Meselson,B.Y.M. and Stahl,W.F. (1958) The replication of DNA in Escherichia coli. *PNAS*, **44**, 671–682.

36. Linn,S. and Arber,W. (1968) Host specificity of DNA produced by Escherichia coli, X. In vitro restriction of phage ED replicative form. *Biochemistry*, **59**, 1300–1306.

37. Arber,W. and Wauters-Willems,D. (1970) Host specificity of DNA produced by Escherichia coli. *Mol. Gen. Genet.*, **108**, 203–217.

38. Arber,W. and Wauters-Willems,D. (1970) Host specificity of DNA produced by Escherichia coli. XII. The two restriction and modification systems of strain 15T-. *Mol. Gen. Genet.*, **108**, 203–217.

39. Meselson,M. and Yuan,R. (1968) DNA restriction enzyme from E.coli. *Nature*, **217**, 1110–1114.

40. Horiuchi,K. and Zinder,D.N. (1972) Cleavage of bacteriophage fl DNA by the restriction enzyme of Escherichia coli B. *PNAS*, **69**, 3220–3224.

41. Smith,H.O. and Wilcox,K.W. (1970) A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. *J. Mol. Biol.*, **51**, 379–391.

42. Willey, Joanne M, Linda Sherwood, Christopher J. Woolverton, Lansing M. Prescott, and J.M.W. (2011) Prescott's microbiology New York: McGraw-Hill.

43. Kauc,L. and Piekarowicz,A. (1978) Purification and properties of a new restriction endonuclease from Haemophilus influenzae Rf. *Eur. J. Biochem.*, **92**, 417–426.

44. De Backer,O. and Colson,C. (1991) Identification of the recognition sequence for the M.StyLTI methyltransferase of Salmonella typhimurium LT7: an asymmetric site typical of type-III enzymes. *Gene*, **97**, 103–107.

45. Mural,R.J., Chesney,R.H., Vapnek,D., Kropf,M.M. and Scott,J.R. (1979) Isolation and characterization of cloned fragments of bacteriophage P1 DNA. *Virology*, **93**, 387–397.

46. Iida,S., Meyer,J., Baechi,B., Stalhammar-carlemalm,M., Schrickel,S., Bickle,T.A. and

Arber,W. (1983) D N A restriction-modification genes of Phage P1 and plasmid pl5B. *J. Mol. Biol.*, **165**, 1–18.

47. Huembelin,M., Surit,B., Rao,D.N., Hornbys,D.P., Eberlejj,H., Pripfl,T., Kenel,S. and Bickle,T.A. (1988) Type III DNA restriction and modfication EcoPl and EcoP15 nucleotide systems sequence of the EcoPl operon , the EcoP15 mod gene and some EcoPl mod mutants. *J. Mol. Biol.*, **200**, 23–29.

48. Ahmad,I., Krishnamurthy,V. and Rao,D.N. (1995) DNA recognition by the EcoP15I and EcoPI modification methyltransferases. *Gene*, **157**, 143–147.

49. Janscak,P., Sandmeier,U., Szczelkun,M.D. and Bickle,T.A. (2001) Subunit assembly and mode of DNA cleavage of the Type III restriction endonucleases EcoP1I and EcoP15I. *J. Mol. Biol.*, **306**, 417–431.

50. Wyszomirski,K.H., Curth,U., Mackeldanz,P., Schutkowski,M., Kru,D.H. and Mo,E. (2011) Type III restriction endonuclease EcoP15I is a heterotrimeric complex containing one Res subunit with several DNA-binding regions and ATPase activity. *Nucleic Acids Res.*, 10.1093/nar/gkr1239.

51. Gupta,Y.K., Yang,L., Chan,S., Samuelson,J.C., Xu,S. and Aggarwal,A.K. (2012) Structural insights into the assembly and shape of type III restriction – modification (R – M) EcoP15I complex by small-angle X-ray scattering. *J. Mol. Biol.*, 10.1016/j.jmb.2012.04.026.

52. Butterer,A., Pernstich,C., Smith,R.M., Sobott,F., Szczelkun,M.D. and Tóth,J. (2014) Type III restriction endonucleases are heterotrimeric: comprising one helicase-nuclease subunit and a dimeric methyltransferase that binds only one specific DNA. *Nucleic Acids Res.*, **42**, 5139–50.

53. Gupta,Y.K., Chan,S.-H., Xu,S. and Aggarwal,A.K. (2015) Structural basis of asymmetric DNA methylation and ATP-triggered long-range diffusion by EcoP15I. *Nat. Commun.*, **6**, 1–10.

54. Hattman,S., Brooks,J.E. and Masurekar,M. (1978) Sequence specificity of the P1 modification methylase (M·EcoP1) and the DNA methylase (M·Ecodam) controlled by the Escherichia coli dam gene. *J. Mol. Biol.*, **126**, 367–380.

55. Rhoads,A. and Au,K.F. (2015) PacBio sequencing and its applications. *Genomics. Proteomics Bioinformatics*, **13**, 278–289.

56. Timinskas,A., Butkus,V. and Janulaitis,A. (1995) Sequence motifs characteristic for DNA [cytosine-N4] and DNA [adenine-N6] methyltransferases . Classification of all DNA methyltransferases. *Gene*, **157**, 3–11.

57. Malone,T., Blumenthal,R.M., Cheng,X. and Structural,W.M.K. (1995) Structure-guided analysis reveals nine sequence motifs conserved among DNA amino-methyl-transferases , and suggests a catalytic mechanism for these enzymes. *J. Mol. Biol.*, **253**, 618–632.

58. Saha,S., Ahmad,I., Reddy,Y. V, Krishnamurthy,V. and Rao,D.N. (1998) Functional analysis of conserved motifs in type III restriction-modification enzymes. *Biol. Chem.*, **379**, 511–517.

59. Gorbalenya,A.E. and Koonin,E. V (1991) Endonuclease (R) subunits of type-I and type -III restriction-modication enzymes contain a helicase-like domain. *FEBS Lett.*, **291**, 277–281.

60. Fairman-williams,M.E., Guenther,U. and Jankowsky,E. (2010) SF1 and SF2 helicases : family matters. *Curr. Opin. Struct. Biol.*, **20**, 313–324.

61. Schwarz,F.W., Toth,J., van Aelst,K., Cui,G., Clausing,S., Szczelkun,M.D. and Seidel,R. (2013) The helicase-like domains of type III restriction enzymes trigger long-range diffusion along DNA. *Science (80-.).*, **340**, 353–356.

62. Singleton,M.R., Dillingham,M.S. and Wigley,D.B. (2007) Structure and mechanism of helicases and nucleic acid translocases. *Annu. Rev. Biochem.*, **76**, 23–50.

63. Mcclelland,S.E. and Szczelkun,M.D. (2004) The Type I and III restriction endonucleases : Structural elements in molecular motors that process DNA. **14**.

64. Aravind,L., Makarova,K.S. and Koonin,E. V (2000) Holliday junction resolvases and related nucleases : identification of new families , phyletic distribution and evolutionary trajectories. *Nucleic Acids Res.*, **28**, 3417–3432.

65. Kovall,R.A. and Matthews,B.W. (1998) Structural, functional, and evolutionary

relationships between lambda-exonuclease and the type II restriction endonucleases. *PNAS*, **95**, 7893–7897.

66. Kovall,R. and Matthews,B.W. (1997) Toroidal structure of lambda-exonuclease. *Science.*, **277**, 1824–1827.

67. Winkler,F.K., Banner,D.W., Oefner,C., Tsernoglou,D., Brown,R.S., Heathman,S.P., Bryan,R.K., Martin,P.D., Petratos,K. and Wilson,K.S. (1993) The crystal structure of EcoRV endonuclease and of its complexes with cognate and non-cognate DNA fragments. *EMBO J.*, **12**, 1781–1795.

68. Galburt,E.A. and Stoddard,B.L. (2002) Catalytic mechanisms of restriction and homing endonucleases. *Biochemistry*, **41**, 13851–13860.

69. Horton,N.C. and Perona,J.J. (2004) DNA cleavage by EcoRV endonuclease: two metal ions in three metal ion binding sites. *Biochemistry*, **43**, 6841–6857.

70. Ramanathan,S.P., Aelst,K. Van, Sears,A., Peakman,L.J., Diffin,F.M., Szczelkun,M.D. and Seidel,R. (2009) Type III restriction enzymes communicate in 1D without looping between their target sites. *PNAS*, **106**, 1748–1753.

71. Merlind,M., Reich,S., Mo,E., Reuter,M. and Kru,D.H. (2001) DNA Cleavage by Type III restriction-modification enzyme EcoP15I is independent of spacer distance between two head to head oriented recognition. *J. Mol. Biol.*, **312**, 687–698.

72. Sears,A., Peakman,L.J., Wilson,G.G. and Szczelkun,M.D. (2005) Characterization of the Type III restriction endonuclease PstII from Providencia stuartii. *Nucleic Acids Res.*, **33**, 4775–4787.

73. Peakman,L.J. and Szczelkun,M.D. (2009) S-Adenosyl homocysteine and DNA ends stimulate promiscuous nuclease activities in the Type III restriction endonuclease EcoPI. *Nucleic Acids Res.*, **37**, 3934–3945.

74. Möncke-buchner,E., Rothenberg,M., Reich,S., Wagenführ,K., Matsumura,H., Terauchi,R., Krüger,D.H. and Reuter,M. (2009) Functional characterization and modulation of the DNA cleavage efficiency of Type III restriction endonuclease EcoP15I in its interaction with two sites in the DNA target. *J. Mol. Biol.*, **387**, 1309–1319.

75. Raghavendra,N.K. and Rao,D.N. (2004) Unidirectional translocation from recognition site and a necessary interaction with DNA end for cleavage by Type III restriction enzyme. *Nucleic Acids Res.*, **32**, 5703–5711.

76. Peakman,L.J., Antognozzi,M., Bickle,T.A., Janscak,P. and Szczelkun,M.D. (2003) S-Adenosyl methionine prevents promiscuous DNA cleavage by the EcoP1I type III restriction enzyme. *J. Mol. Biol.*, **333**, 321–335.

77. Smith,R.M., Josephsen,J. and Szczelkun,M.D. (2009) The single polypeptide restriction – modification enzyme LlaGI is a self-contained molecular motor that translocates DNA loops. *Nucleic Acids Res.*, **37**, 7219–7230.

78. Nidhanapati,R., Hyderabad,T., Bheemanaik,S. and Sagar,D. (2012) Mechanistic insights into type III restriction enzymes. *Fron. in Biosci.,* 10.2741/3975.

79. Dryden,D.T.F., Edwardson,J.M. and Henderson,R.M. (2011) DNA translocation by type III restriction enzymes : a comparison of current models of their operation derived from ensemble and single-molecule measurements. *Nucleic Acids Res.*, 10.1093/nar/gkq1285.

80. Crampton,N., Yokokawa,M., Dryden,D.T.F., Edwardson,J.M., Rao,D.N., Takeyasu,K., Yoshimura,S.H. and Henderson,R.M. (2007) Fast-scan atomic force microscopy reveals that the type III restriction enzyme EcoP15I is capable of DNA translocation and looping. *PNAS*, **104**, 12755–12760.

81. Crampton,N., Roes,S., Dryden,D.T.F., Rao,D.N., Edwardson,J.M. and Henderson,R.M. (2007) DNA looping and translocation provide an optimal cleavage mechanism for the type III restriction enzymes. *EMBO J.*, **26**, 3815–3825.

82. Bandyopadhyay,P.K. (1988) Model for how type I restriction enzymes select cleavage sites in DNA. **85**, 4677–4681.

83. Murray,N.E. (2000) Type I restriction systems: sophisticated molecular machines (a Legacy of Bertani and Weigle). *Microbiol. Mol. Biol. Rev.*, **64**, 412–434.

84. Haberman,A. (1974) The bacteriophage P1 restriction endonuclease. *J. Mol. Biol.*, **89**, 545–563.

CHAPTER 2

Cloning, Purification and Characterization of EcoP1I

Chapter 2

Cloning, Purification and Characterization of EcoP1I

1. Introduction

Type III RM enzymes are heterotrimeric oligomers made of two subunits - methyltransferase (Mod) and restriction endonuclease (Res). Mod facilitates binding of the enzyme to specific target sequence, and carries out sequence specific methylation of adenine at N6 position (1, 2). In complex with Mod, Res (a fusion of ATPase and endonuclease domains connected by a linker) cleaves unmodified DNA having two inversely oriented asymmetric target sites (3–5). The characteristics of a typical asymmetric target site recognized by Type III RM enzymes are described in Chapter 1. Single molecule studies such as magnetic tweezers assay combined with TIRF have suggested that upon recognition of unmodified target site, hydrolysis of ATP allows the enzyme to undergo a conformational change in order to execute one-dimensional motion on the DNA track (6, 7). The diffusing enzyme when encounters another juxtaposed site bound enzyme, the physical interaction between the two leads to activation of the endonuclease to bring about dsDNA cleavage (8, 9). Interestingly, unlike most other ATPase motors that consume approximately one ATP per base pair, the ATPase of the Type III enzyme hydrolyses approximately 30 nucleotides irrespective of the distance travelled (7). These energy efficient machines bring about long range communication between two sites separated by as many as few thousands of base pairs (9). Type III RM enzymes hence present a unique mode of long range communication, i.e. involving an ATP induced molecular switch communicating over long distances via one dimensional diffusive sliding.

A partial structure of EcoP15I bound to a DNA substrate mimic was published recently (10). This structure lacks information about the endonuclease domain due to weak electron densities, mostly because of flexibility associated with this region. Due to lack of information on interaction of linker and endonuclease with DNA, a complete picture of cleavage competent EcoP1I-DNA complex cannot be understood. Although biochemical, biophysical and crystallographic studies aimed at understanding the working of these machines have provided significant information till date, mechanistic details of how these enzymes nucleolytically cleave dsDNA are still lacking. For example, how does recognition of target site by Mod activates Res to hydrolyse ATP and

nucleolytically cleave DNA downstream of the two target sites; how does the Mod_2Res assembly switch from a loading mode to diffusive mode after hydrolysis of ATP; does Mod loosen the grip over target site in order to facilitate bidirectional diffusion on DNA track; does slackening of Mod-DNA grip result in rearrangements of Res-DNA interactions; given that there exists only one ATPase domain which serves as a motor, how does the macromolecular assembly undergo bidirectional diffusion.

To better understand the molecular mechanism and gain structural insights into their working, we initiated crystallographic studies of EcoP1I, a prototype of Type III R-M enzymes. Crystallographic studies require a large amount of homogenous and pure protein. We started by cloning EcoP1I in vectors suitable for overexpression of protein. In addition to the operon encoding the wild type EcoP1I, we also cloned *mod* and *res* separately, the gene fragment encoding the ATPase domain, nuclease domain and full-length operon of EcoP1I. To avoid any aberrant cleavage during crystallisation of the EcoP1I-DNA complex, a point mutant EcoP1I (E916A) and a double mutant EcoP1I (D898A-E916A) lacking nuclease activities were generated. Previous studies show that these mutants have no endonuclease activity (11). Subsequent to cloning, a new protocol for purification of EcoP1I in large scale primarily for crystallographic studies was standardised.

The enzymatic activities of EcoP1I purified by this new method were checked using DNA binding, ATP hydrolysis and nuclease assays. These studies were also important to design crystallisation experiments described in Chapter 5. We studied the DNA binding affinity of EcoP1I for different types of substrates, such as long linear, supercoiled (SC) and short DNA. Earlier studies reported differential effects of potassium and sodium ions on cleavage by EcoP1I (12), hence effect of the two cations on DNA binding was also examined. These studies were followed with measurement of the active fraction of enzyme in the purified sample using ATP hydrolysis kinetics of EcoP1I. Next we examined the ability of the enzyme to cleave DNA. Cleavage assays were done with supercoiled as well as linear DNA substrates. Effect of orientations of target sites viz, Head-to-Head, Tai-to-Tail and Head-to-Tail, on DNA cleavage was also examined. The assays confirmed that EcoP1I purified by new protocol was enzymatically active. This chapter describes the methods and discusses the results of the above-mentioned assays.

2. Materials and Methods

2.1 Cloning

EcoP1I operon, *mod*, *res*, and gene-fragments encoding the ATPase domain and the nuclease domain were amplified from recombinant plasmid construct pET11b-EcoP1I, which was provided by Prof. D.N.Rao, IISc, Bangalore India (13). The boundaries of ATPase and nuclease domains were delineated based on bioinformatics and biochemical data (14). Bioinformatics strategies used are described in Chapter 4. Details of different inserts used in the study are given in Figure 2.1. The amplified products were subcloned into the vectors pHis17 (15) with a C-terminal hexahistidine tag and pRSF-1b with an N-terminal hexahistidine tag, respectively. pHis17 has ampicillin selection marker and pRSF-1b has kanamycin selection marker. Primers were purchased from Integrated DNA Technologies (IDT), USA and Sigma-Aldrich®, USA. Table 2.1 lists the sequences and lengths of primers used for cloning various subunits and domains of EcoP1I. For Polymerase Chain Reaction (PCR) amplifications, PfuTurbo® was purchased from Agilient and Accuprime™ Pfx polymerase was purchased from Thermo Fischer Scientific. Cloning was done using either digestion-ligation protocol or restriction free (RF) cloning (16). Restriction enzymes and T4 DNA ligase were purchased from New England BioLabs® Inc (NEB).

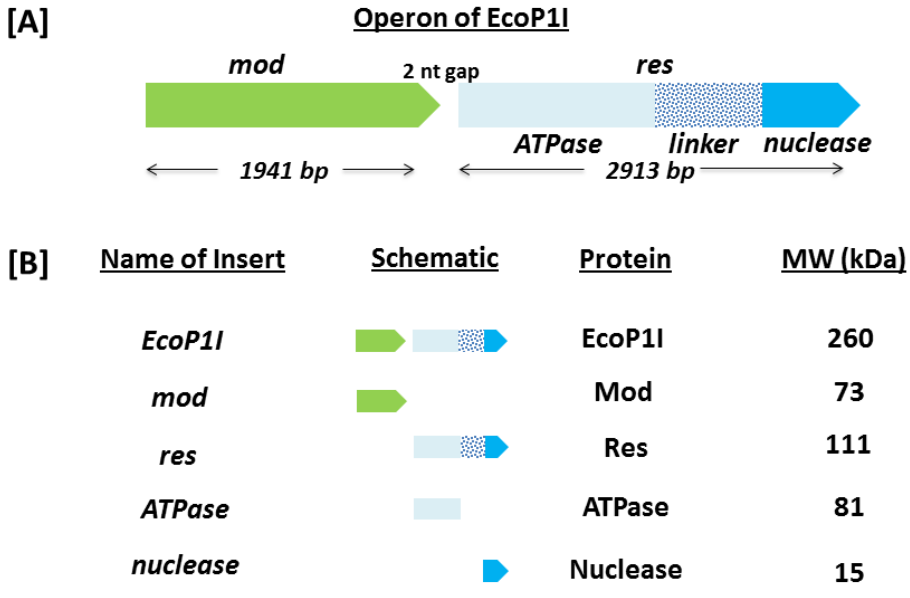


Figure 2.1: Cloning strategy. A] The EcoP1I operon (4856 bp) has *mod* (green arrow) upstream of *res* (blue arrow). *res* is a fusion of ATPase (light blue rectangle) and nuclease (cyan arrow). The linker between ATPase and nuclease is shown as dotted blue rectangle. B] Names of different inserts along with their corresponding translation product and molecular weight.

Table 2.1: Primers used for cloning of EcoP1I

Sr No	Name	Sequence (5' -> 3')
1	EP1RPHISF	GTTTAACTTTAAGAAGGAGATATACATATGTCAAAAGGGTTCACATTCGAAAAG
2	EP1RPHISR1	CTTTTAATGATGATGATGATGATGGGATCCTGGTAATGCGCTCTTGATTAAC
3	EP1RPHISR2	CTTTTAATGATGATGATGATGATGGGATCCttaTGGTAATGCGCTCTTGATTAAC
4	EP1MPRSFRFF	AACTTTAAGAAGGAGATATACCATGGATGAAAAAAGAAACGATTTTTTCCG
5	EP1MPRSFRFR	CGCAGCAGCGGTTTCTTTACCGAGCTCTTAGTTCCTTACCACTAAATCTAAC
6	EP1HPRSFF	AACTTTAAGAAGGAGATATACCATGGATGATGTCAAAAGGGTTCACATTC
7	EP1HPRSFR	CGCAGCAGCGGTTTCTTTACCCTCGAGTTATGATTTGGTAAATCGTTCTG
8	EP1RPRSFR	CGCAGCAGCGGTTTCTTTACCCTCGAGTTATGGTAATGCGCTCTTGATTAAC
9	EP1MPHISF	GTTTAACTTTAAGAAGGAGATATACATATGAAAAAAGAAACGATTTTTTCCG
10	EP1MPHISR1	CTTTTAATGATGATGATGATGATGGGATCCTTAGTTCCTTACCACTAAATC
11	EP1MPHISR2	CTTTTAATGATGATGATGATGATGGGATCCGTTCCTTACCACTAAATC
12	EP1MNdeSi	CATAAGCATTAACCATGTGCTTCAATACTTC
13	Ecp1MpRSFTEV1	GATATACCATGGCACATCACCACCACCATCACGAGAACCTGTACTTCCAAGGCATGAAAAAAGAAACGATTTTTTCCG
14	EP1E916AF	GGTGATTATCTTAACTTTATCATTGCAACAAAAAACGTAGATAGTAAGGATAG
15	EP1E916AR	CTATCCTTACTATCTACGTTTTTTGTTGCAATGATAAAGTTAAGATAATCACC
16	Ecp1D898A1	GCTGGTGGATACACTTACTCACCCGCTTTTGCTTATGTTGTAAAAACAGCAGAAGG
17	Ecp1D898A2	CCTTCTGCTGTTTTTACAACATAAGCAAAAGCGGGTGAGTAAGTGTATCCACCAGC
18	P1NucC1-F	GTTTAACTTTAAGAAGGAGATATACATATGTCAAAAGCACCATTAG
19	P1NucC2-F	GTTTAACTTTAAGAAGGAGATATACATATGTTTGAAGAAGTTTTTTATG
20	PHISMCRCF	GATGAACTTTAAGAAGGAGATATACATATGTCAAAAGCACCATTAG
21	PHISMCRCR	CGAACTTTAAGAAGGAGATATACATATGTTTGAAGAAGTTTTTTATG

2.1.1 Digestion-ligation protocol:

The gene of interest was amplified using PCR. The primers for PCR amplification were designed to introduce NdeI and BamHI restriction sites upstream and downstream of the genes, respectively. This was designed to facilitate ligation into pHis17 vector. *mod* has an NdeI site 321 bp downstream of start codon. This site was silenced without affecting the amino acid sequence by amplifying a shorter insert using EP1MPHISF as forward primer and EP1MNdeSi as reverse primer (Table 2.1). EP1MNdeSi had fourth adenine of NdeI target site (5' CAT**A**TG 3') replaced by a cytosine. This amplicon was subsequently used as forward primer to amplify inserts suitable for cloning in pHis17. To clone the gene into pRSF-1b, NcoI and XhoI restriction sites were introduced. The amplified insert and the template plasmids were digested completely with the appropriate pair of restriction enzymes to obtain inserts and vectors with complementary sticky ends. The ligation of insert and template plasmid was carried out using T4 DNA Ligase based on the protocol by NEB®. Electrocompetent *E.coli* NEB®Turbo cells were transformed with the recombinant plasmids by the method of electroporation using Gene Pulser Xcell™ Electroporation Systems from Bio-Rad. A Gene Pulser® electroporation cuvette with 0.2 cm gap was used to electroporate the cells. A negative control of restriction-digested vector without insert was used to transform NEB®Turbo cells. The transformed cells were plated on agar plates containing lysogeny broth (LB) medium and appropriate antibiotic. The LB agar plates were incubated at 37° C for 12-14 hours. The success of digestion-ligation experiment was confirmed by comparing the number of colonies on test plates with that on negative control. Isolated colonies from the test plates were then picked and inoculated in LB supplemented with appropriate antibiotic. The cells grew under continuous shaking at 37° C for 7-8 hours. Subsequently, the cells were pelleted down and plasmids were recovered by using plasmid mini-prep kit purchased from Qiagen. The purified plasmids were then digested with appropriate restriction enzymes to check for the presence of the insert. Positive clones were confirmed only after sequencing the entire insert.

2.1.2 Restriction Free (RF) cloning:

The primers for restriction free cloning were designed according to guidelines given by Ent and Loewe (16). The gene of interest was amplified using two rounds of PCR. In the first round the short synthetic primers were used for amplification of the gene to be inserted. In the second round of PCR, the amplified insert was used as primers and the vector plasmid was used as template to obtain a plasmid with the insert. Subsequent addition of the restriction enzyme DpnI degraded the methylated template plasmid. This recombinant plasmid was then transformed into electrocompetent *E.coli* NEB®Turbo cells by the method of electroporation. A negative control of template plasmid subjected to a round of PCR without the primer (the insert in this case) was also transformed into another vial of NEB®Turbo cells. Subsequent steps for obtaining and confirming the clone were done as described in section 2.1.1.

2.1.3 Site directed mutagenesis:

As a first step, a single point mutation E916A was introduced. The recombinant plasmid containing EcoP1I(E916A) was used as template to introduce another single point mutation D898A to get a double mutant EcoP1I(D898A,E916A). These single mutations were performed using the QuickChange® protocol of Stratagene. The primers bearing mutations were carefully designed such that the Tm was greater than or equal to 78°C and had at least 40% GC content. The Tm was calculated according to the guidelines given in QuickChange® mutagenesis manual. 12-18 cycles of PCR were performed for single amino acid mutation. The template plasmid was then digested by incubation with DpnI for 2 hours at 37°C. The recombinant plasmid thus obtained was transformed into electrocompetent *E. coli* NEB®Turbo cells by electroporation. A negative control of template plasmid subjected to a round of PCR without the primer (the insert in this case) was also transformed into another vial of NEB®Turbo cells. Subsequent steps for obtaining and confirming the clone were done as described in section 2.1.1.

2.1.4 Sequencing of EcoP1I operon:

The length of EcoP1I operon (4856 bp) is much longer than the sequence read out from a Sanger sequencing reaction (17), which in general is approximately 800 bp. Consequently, T7 promoter, T7 terminator and five other primers complementary to intermittent regions of the genes (Table 2.2) were used to sequence the entire operon.

Table 2.2: Primers for sequencing EcoP1I operon

Sr No	Name	Sequence (5' -> 3')
1	T7 Promoter	TAATACGACTCACTATAGGG
2	S1	CAAGCAGTCATAGTGCATGG
3	S2	GACTTTTTTGCTGGCTCTGG
4	S3	GAAAGCAGGTGTCGATGCGG
5	S4	GTTAAAGGTATTGATGCGT
6	S5	CAAATGGACGTTGCGTGAGG
7	T7 Terminator	GCTAGTTATTGCTCAGCGG

2.2 Expression of EcoP1I

After sequencing, positive recombinant plasmids were subjected to expression of the cloned inserts. The recombinant pHis17 plasmids containing the gene of interest were transformed in *E. coli* BL21(AI). This strain has T7 polymerase gene as a chromosomal insertion and the regulation of its transcription is under tight control of the araBAD operon, which is induced only when L arabinose is supplied in the media. Alternatively, we generated recombinant pRSF-1b plasmids containing our gene of interest. pRSF-1b has T7 promoter and lac operator to control transcription of gene. It has kanamycin resistance marker. The recombinant pRSF-1b plasmids containing the gene of interest were transformed in *E. coli* BL21(DE3). This system is Isopropyl β-D-1-thiogalactopyranoside (IPTG) inducible.

The transformed cells were plated on LB agar plates with a suitable antibiotic. After incubation of the plate at 37°C for 12-14 hours, the colonies were inoculated in LB media supplemented with appropriate antibiotic. The cultures were incubated at 37°C with constant shaking (150 RPM) using Forma Orbital Shaker (Thermo Fisher Scientific). As the culture reached an OD of 0.5, it was induced with 2 g/L L-arabinose for BL21(AI) and 1 mM IPTG for BL21(DE3). Initially, the expression was checked by shaking the cultures continuously at 37°C for 3 hours, at 25°C for 5 hours and at 18°C overnight. The amount of inducer was also optimised to maximise the yield of soluble protein. The cultures were then pelleted down and resuspended in chilled lysis buffer. Cell pellet from 1 mL culture was resuspended in 100 µL lysis buffer (50 mM Tris-Cl (pH 8), 500 mM NaCl, 10 mM $MgCl_2$, 1 mM EDTA). The cells were sonicated on ice for 20 seconds with 50% amplitude and 5 seconds ON, 5 seconds OFF cycle using VibraCell™. The lysate was then centrifuged at 4°C at 15,000 RPM for 20 mins on a table-top microcentrifuge. After centrifugation, the supernatant was separated from the pellet. This pellet was resuspended in 5X Tris-glycine-SDS buffer and sonicated as mentioned before. 1 volume of SDS gel loading dye was added to the sonicated pellet and supernatant containing soluble fraction. The samples were heated at 99°C for 10 minutes and centrifuged thoroughly for 15 minutes. The samples were then loaded on 10% SDS-PAGE gel and electrophoresed along with marker proteins to test the expression of cloned genes. After the expression of protein was confirmed, the solubility of protein was checked by comparing the amount of the over-expressed protein in supernatant versus lysate pellet.

2.3 Purification of EcoP1I

Once the expression of protein was standardized, we performed large-scale purification to achieve sufficient quantity of protein suitable for crystallization experiments. The recipes of various buffers used in purification of EcoP1I are summarized below (Table 2.3).

Table 2.3: Compositions of buffers used for purification of EcoP1I

	Tris pH 8.0 (mM)	NaCl (mM)	Glycerol (%)	MgCl2 (mM)	Imidazole (mM)	$(NH4)_2SO_4$ (M)	KCl (mM)	EDTA (mM)	DTT (mM)
Lysis Buffer	50	150	10	10				1	
Buffer A	50	500			25				
Buffer B	50	500			500				
B0	50	0						1	1
B50	50	50						1	1
B1000	50	1000						1	1
B50 + $(NH4)_2SO_4$	50	50				2		1	1
Buffer C	50	100		10				1	1
Buffer D	50			10			50	1	1

2.3.1 Purification of Mod$^{C\text{-}His}$, Res$^{C\text{-}His}$, ATPase$^{C\text{-}His}$

A 6 L culture pellet was resuspended in 200 mL of lysis buffer A. The resuspended pellet was sonicated on ice for 3 minutes with 70% amplitude and 10 seconds ON, 59 seconds OFF cycle using VibraCell™. The resulting lysate was spun at 37000 RPM at 4°C for 45 minutes using Optima™ XE ultracentrifuge (Beckman Coulter Life Sciences). The supernatant, thus obtained, was loaded on 5 mL HisTrap HP Ni-NTA column (GE Healthcare) pre-equilibrated with buffer A. After loading, the column was again washed with buffer A to remove any unbound proteins. The protein was then eluted (fraction size 5 mL) with a stepwise gradient of imidazole by going from 5% to 100% buffer B. For Mod$^{C\text{-}His}$, fractions containing protein were pooled and concentrated using Vivaspin®2 (GE Healthcare) protein concentrator (MWCO 30 kDa). After concentration, 500 µL of the protein was injected into a Superose™ 6 10/300 GL gel filtration column (GE Healthcare) for size exclusion chromatography (SEC). After gel filtration, the protein sample was further concentrated, aliquots were flash frozen and stored at -80°C.

2.3.2 Purification of EcoP1I (E916A) $^{C\text{-}His}$ *and EcoP1I (D898A, E916A)* $^{C\text{-}His}$

The downstream processing of 6 L culture until affinity chromatography using Ni-NTA column was executed as described in section 2.3.1. After affinity chromatography, fractions containing protein were pooled and dialyzed against 2 L B50 for 2 hours. The dialysate was spun at 18000 RPM for 20 minutes at 4°C using Avanti J-26XP (Beckman Coulter Life Sciences). After the spin, the sample was filtered through a 0.45 mm filter and loaded on pre-equilibrated 8 mL Mono Q™ 10/100 GL column (GE Healthcare). 2ml fractions were collected in 20 column volumes over a linear gradient of 0% to 50% buffer using B50 and B1000. Fractions containing protein were pooled and concentrated in Vivaspin®2 (GE Healthcare) protein concentrator (MWCO 30 kDa). The 500 µL of concentrated protein was injected into gel filtration column Superdex 200 10/300 GL (GE Healthcare) for SEC. After gel filtration, the protein sample was further concentrated, aliquots were flash frozen and stored at -80°C.

2.3.3 Purification of EcoP1I$^{N\text{-}His}$

The purification scheme of full length EcoP1I$^{N\text{-}His}$ was similar to that of EcoP1I (E916A) $^{C\text{-}His}$ and EcoP1I (D898A, E916A) $^{C\text{-}His}$. However, the protein sample obtained after anion exchange chromatography using the Mono Q™ column was not pure enough to load on gel filtration column. To get rid of impurities after ion exchange chromatography, an additional purification step of hydrophobic interaction chromatography (HIC) using HiTrap Phenyl HP Low Substitution (GE Healthcare) was included. Subsequent to ion exchange chromatography, fractions containing EcoP1I were pooled and mixed with 1.5 volumes of B50 + ammonium sulfate (See Table 2.3). The mixture was loaded on pre-equilibrated 5 mL HiTrap Phenyl HP Low Substitution column (GE Healthcare). The protein was eluted (fraction size 5 mL) with a linear gradient from 100 % to 0 % of ammonium sulphate in buffer B50. After HIC purification scheme similar to that of EcoP1I (E916A) $^{C\text{-}His}$ and EcoP1I (D898A, E916A) $^{C\text{-}His}$ was followed.

2.3.4 Purification of untagged EcoP1I

A 6 L culture pellet was resuspended in 200 mL lysis buffer. The resuspended pellet was sonicated as described in section 2.3.1. The resulting lysate was spun at 37000

RPM at 4°C for 45 minutes using Optima™ XE ultracentrifuge (Beckman Coulter Life Sciences). The resulting supernatant was saturated with 25% ammonium sulphate while constant stirring at 4 °C over a period of 10-15 minutes followed by a spin at 18000 RPM for 20 minutes at 4°C using Avanti J-26XP (Beckman Coulter Life Sciences). The supernatant from this spin was again saturated with 75% ammonium sulphate while constant stirring at 4 °C over a period of 10-15 minutes. The mix was spun at 18000 RPM for 20 minutes at 4°C. The supernatant was decanted and the pellet was resuspended in 300 mL of B0. The resuspended pellet was loaded on pre-equilibrated 5 mL HiTrap Heparin HP column (GE Healthcare). After loading, the column was washed with B50 to remove any unbound proteins. The protein was eluted with a linear gradient of buffers B50 to B1000 over 20 CV. The fractions containing protein were pooled and dialyzed against 2 L B50 for 2 hours. After dialysis, the purification scheme similar to that of EcoP1I$^{N\text{-}His}$ was followed.

2.4 Substrate generation for DNA binding, ATPase and cleavage Assays

A specific linear DNA fragment was obtained by PCR amplification of *mod* using primers EPIMPHISF and EP1MPHISR1 from pHis17-EcoP1I. A non-specific (without target sites) linear DNA fragment was amplified using primers MCRCPHISF and MCRCPHISR from pHis17-McrC (kind gift from Neha Nirwan). The specific DNA (1941 bp) had a single target site 891 bp from the upstream end and 1045 bp from the downstream end. This DNA will be referred to as 891/1045_P1.

To generate short DNA substrates, oligomers were purchased from Integrated DNA Technologies, USA, and Sigma-Aldrich, USA. dsDNA were made by annealing the single-strands and further purified using an 8 mL MonoQ™ 10/100 GL column (GE Healthcare). The resulting duplexes were washed thoroughly with MilliQ to remove any salt and concentrated using Vivaspin®2 (GE Healthcare) concentrator (MWCO 3 kDa). The concentrated oligomer was stored in MilliQ at -80°C until further use. A list of dsDNA used for various biochemical assays described below are listed in Table 2.4. The F and R pairs were annealed to get a duplex DNA.

Table 2.4: Sequences of oligomers used for biochemical assays with EcoP1I

Name	Sequence (5' -> 3')
32/32_P1_F	CATGACTGAGTCAATCGGATCGTAGACGTACTAGACCTATCCTGTATGCTACGTATTCGTATCGTGAGC
32/32_P1_R	GCTCACGATACGAATACGTAGCATACAGGATAGGTCTAGTACGTCTACGATCCGATTGACTCAGTCATG
30/32_P1_F	TGACTGAGTCAATCGGATCGTAGACGTACTAGACCTATCCTGTATGCTACGTATTCGTATCGTGAGC
30/32_P1_R	GCTCACGATACGAATACGTAGCATACAGGATAGGTCTAGTACGTCTACGATCCGATTGACTCAGTCA
15/32_P1_F	GATCGTAGACGTACTAGACCTATCCTGTATGCTACGTATTCGTATCGTGAGC
15/32_P1_R	GCTCACGATACGAATACGTAGCATACAGGATAGGTCTAGTACGTCTACGATC
10/32_P1_F	TAGACGTACTAGACCTATCCTGTATGCTACGTATTCGTATCGTGAGC
10/32_P1_R	GCTCACGATACGAATACGTAGCATACAGGATAGGTCTAGTACGTCTA
5/32_P1_F	GTACTAGACCTATCCTGTATGCTACGTATTCGTATCGTGAGC
5/32_P1_R	GCTCACGATACGAATACGTAGCATACAGGATAGGTCTAGTAC
2/32_P1_F	CTAGACCTATCCTGTATGCTACGTATTCGTATCGTGAGC
2/32_P1_R	GCTCACGATACGAATACGTAGCATACAGGATAGGTCTAG
NS_P1_F	GGCTCACGCTACGACTACGTAGCATACAGGTGACCGTCTAGTACGTCTACGTTCCGAGTGACTCAGTCA
NS_P1_R	TGACTGAGTCACTCGGAACGTAGACGTACTAGACGGTCACCTGTATGCTACGTAGTCGTAGCGTGAGCC

To generate supercoiled (SC) substrate for binding and cleavage studies pUC18, which has 3 target sites for EcoP1I, was selected. Arrangement of these sites is shown in Figure 2.2. The specific site 5’ AGACC 3’ (shown as a red triangle in Figure 2.2) was changed to non-specific site 5’ AGCCC 3’ by site directed mutagenesis to create a plasmid with two target sites oriented Head-to-Head. In addition, we

introduced an XhoI site by site directed mutagenesis between the tails of these sites as shown in Figure 2.2. To generate a Head-to-Tail substrate, we flipped the site shown in green. The resultant plasmids were then designated as pHtH (Head-to-Head) and pHtT (Head-to-Tail). The primers used for these mutagenesis experiments are listed in Table 2.5. Using the above plasmids digested with appropriate restriction enzymes we could also obtain linear DNA substrates with different site orientations (see Figure 2.2). The digested DNA fragments were excised from agarose gel and purified using QIAquick® Gel Extraction Kit (Qiagen) for further assays with EcoP1I.

Table 2.5: Primers used for pUC18 mutagenesis

Sr No	Name	Sequence (5' -> 3')
1	PUC18_MidSil_F	CTAAAGTATATATGAGTAAACTTGGGCTGACAGTTACCAATGCTTAATCAG
2	PUC18_MidSil_R	CTGATTAAGCATTGGTAACTGTCAGCCCAAGTTTACTCATATATACTTTAG
3	PUC18_XhoI_F	CTAAAGTATACTCGAGTAAACTTGGGCTGACAGTTACCAATGCTTAATCAG
4	PUC18_XhoI_R	CTGATTAAGCATTGGTAACTGTCAGCCCAAGTTTACTCGAGTATACTTTAG
5	PUC_HtT_F1	GATCCTTTGATCTTTTCTACGTCTGGGACGCTCAGTGGAACGAAAACTC
6	PUC_HtT_R1	GAGTTTTCGTTCCACTGAGCGTCCCAGACGTAGAAAAGATCAAAGGATC

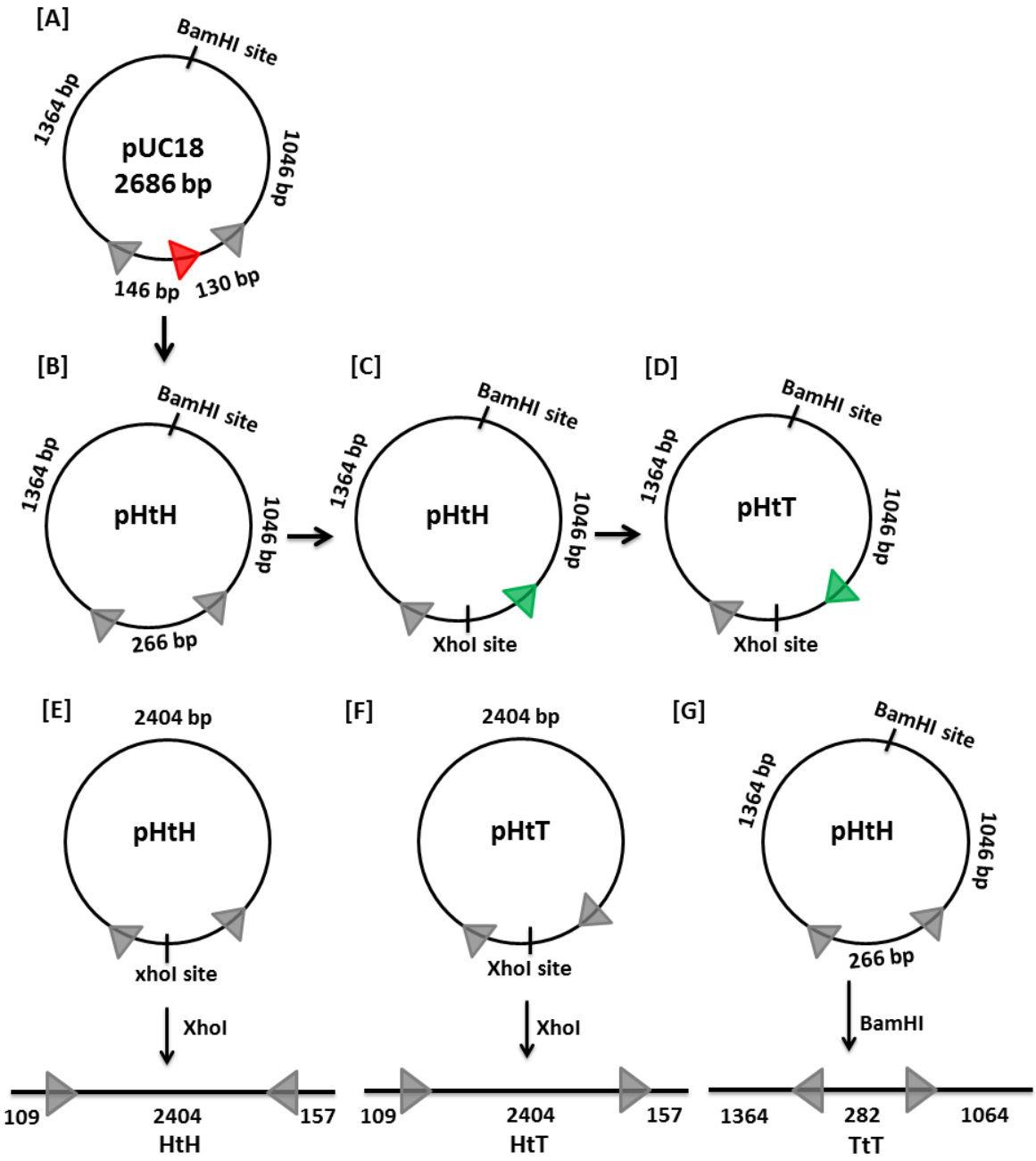


Figure 2.2: Substrate generation for cleavage assays. A] Plasmid vector pUC18 (2686 bp) has 3 target sites of EcoP1I (5' AGACC 3') shown as triangles. XhoI/BamHI restriction site are also illustrated; B] Supercoiled HtH pUC18 substrate was generated by silencing one of the sites (shown as red triangle in A). C] An XhoI site was introduced between two tails of target sequences. D] Further one of the target sequences (shown as green triangle) was flipped to generate supercoiled HtT pUC18**.** E] pHtH cleaved with XhoI site gave linear HtH substrate. F] pHtT cleaved with XhoI gave linear HtT substrate. G] pHtH cleaved with BamHI gave linear TtT substrate.

2.5 Electrophoretic Mobility Shift Assay (EMSA)

For long DNA substrates, the protein and DNA mix was incubated for 40 minutes at 25° C in buffer C+ or D+ (Buffer C/D supplemented with 100 µg/ml BSA). Subsequent to incubation, a native stop buffer (10 mM Tris, pH8, 100 mM EDTA, 40% w/v sucrose, 0.03% bromophenol blue, 0.03% xylene cyanol) was added and immediately loaded on a 1% agarose gel and electrophoresed in 1X TAE (Tris-Acetate-EDTA). The gels were scanned using GeneSnap scanner from Syngene. For short DNA substrates, the protein and DNA mix was incubated for 40 minutes at 25° C in buffer D+. Subsequent to the incubation, a native stop buffer (10 mM Tris, pH8, 60 mM EDTA, 60% glycerol, 0.03% bromophenol blue, 0.03% xylene cyanol) was added and immediately loaded on a 5% native PAGE gel [5% acrylamide:bis-acrylamide(29:1), 1X TBE]. The gels were run at 4°C in 1X TBE (Tris-Borate-EDTA), stained with ethidium bromide and scanned using Typhoon TRIO+ variable mode imager (GE Healthcare).

2.6 Cleavage assays

Cleavage assays were carried out at 25°C in buffer D+. DNA and protein were incubated for 40 minutes. Subsequent to the incubation, 1 mM ATP (Sigma-Aldrich) was added and incubated further for 15 minutes. The reactions were stopped by adding 0.5 volumes of a stop buffer (10 mM Tris, pH8, 100 mM EDTA, 40% w/v sucrose, 0.025% SDS, 0.03% bromophenol blue, 0.03% xylene cyanol) and loaded on 1% agarose gel and electrophoresed. The gels were scanned using GeneSnap scanner from Syngene.

2.7 ATPase assays

The reactions for the ATPase assay were done similar to cleavage assay in buffer D+. The reactions were stopped at fixed time points by addition of 100 mM EDTA. ATPase activity was checked using malachite green (18, 19). Malachite green solution was prepared by mixing 0.044 g malachite green to 36 mL 3 N H_2SO_4. For the detection of inorganic phosphate, the dye solution was prepared freshly by mixing 800 µL of malachite green solution, 200 µL 7.5% ammonium molybdate and 16 µL of 11% Tween20. For every 20 µL reaction, 50 µL malachite green dye was added. The absorbance at 630 nm was recorded after 15 minutes using Varioskan

(Thermo Fischer Scientific). The inorganic phosphate released was calculated from a standard phosphate curve using Na_2HPO_4 plotted for every assay. For data analysis, the data was fit using CurveExpert Professional (Version 2.4.0) and initial rates of reaction were calculated by determining the slope of initial linear portion of the time course.

3. Results

3.1 Expression of EcoP1I

Each clone was checked for expression of protein under various conditions detailed in section 2.2. It was observed that these proteins expressed only from those cultures obtained by inoculating freshly grown colonies. The amount of protein expressed from various constructs is shown in Figure 2.3. After the initial expression check, expression conditions were standardized to maximise soluble protein. All constructs yielded good amount of soluble proteins when the cultures were grown at 18°C post-induction. Mod$^{C\text{-}His}$ could be expressed at 37°C (induction with 2 g/L of L arabinose); however solubility increased as the cultures were shifted to 18°C. Similarly ATPase$^{C\text{-}His}$ and Res$^{C\text{-}His}$ were poorly soluble at 37°C and 25°C while solubility marginally increased at 18°C upon induction with 1 g/L of L arabinose. Nuclease$^{C\text{-}His}$ was found to be expressed in good amount, however the protein was always found insoluble (data not shown).

Along with expressing different subunits and domains of EcoP1I, the recombinant plasmids containing operon was also subjected to expression trials (See Figure 2.1 for details). This involved co-expression of Mod$^{C\text{-}His}$ and Res$^{C\text{-}His}$ under single promoter. Co-expression of Res and Mod from four different constructs, viz. EcoP1I(E916A)$^{C\text{-}His}$, EcoP1I(D898A,E916A)$^{C\text{-}His}$, EcoP1I$^{N\text{-}His}$ and untagged EcoP1I, at 18°C (induction with 1 g/L of L arabinose for BL21(AI) and 0.5 mM IPTG for BL21(DE3)) yielded soluble protein complex.

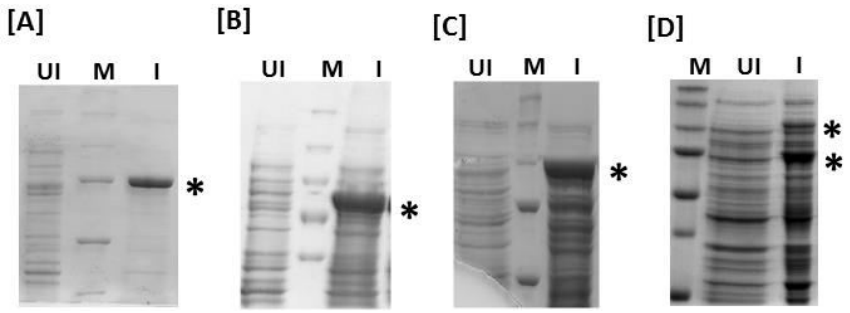


Figure 2.3: Expression of various constructs. The expressed protein is highlighted by asterisk. UI: Uninduced, I: Induced, M: Marker (Precision Plus™ All blue ladder BioRad). A] Mod$^{C\text{-}His}$; B] ATPase$^{C\text{-}His}$; C] Res$^{C\text{-}His}$; D] EcoP1I.

3.2 Protein purification

After optimizing expression conditions for a given construct, large scale purification of protein was initiated. Figure 2.4 shows the purification schemes of various constructs. Mod$^{C\text{-}His}$ was purified in one-step by affinity chromatography. The protein was sufficiently pure and could be directly subjected to SEC (Figure 2.4, A). Res$^{C\text{-}His}$ or ATPase$^{C\text{-}His}$ though expressed well (Figure 2.3), were found to be predominantly insoluble. Attempts to purify them by affinity chromatography using HisTrap HP Ni-NTA column (GE Healthcare) yielded very little protein (Figure 2.4, B and C). ATPase$^{C\text{-}His}$ could be purified in small amounts using Ni-NTA affinity chromatography (Figure 2.4, B). However, this protein was found to be aggregated using SEC (data not shown). Consequently, Res$^{C\text{-}His}$ and ATPase$^{C\text{-}His}$ were not included in any subsequent studies.

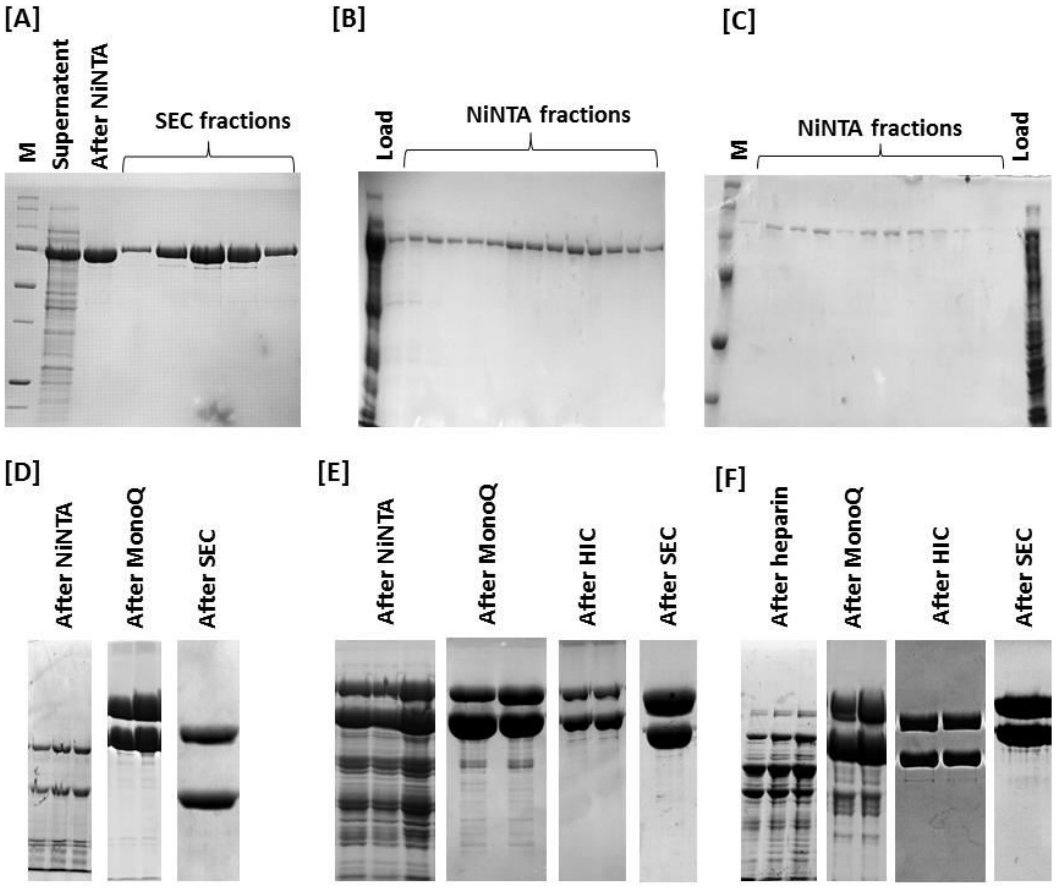


Figure 2.4: Purification scheme of various constructs. A] Mod$^{C\text{-}His}$; B] ATPase$^{C\text{-}His}$; C] Res$^{C\text{-}His}$; D] EcoP1I(E916A)$^{C\text{-}His}$; E] EcoP1I$^{N\text{-}His}$; F] Untagged EcoP1I. M: Marker (Precision Plus™ All blue ladder BioRad).

EcoP1I (E916A)$^{C\text{-}His}$ and EcoP1I (D898A, E916A)$^{C\text{-}His}$ were purified by including an ion exchange chromatography step after Ni-NTA affinity chromatography. The protein thus obtained was pure enough to be used for SEC subsequently (Figure 2.4, D). EcoP1I$^{N\text{-}His}$ was purified in a similar manner; however the samples contained impurities after ion exchange chromatography. To purify the protein further an additional step of HIC was included. After HIC the protein sample attained >98 % purity (Figure 2.4E).

The untagged EcoP1I was purified from crude lysate by using heparin affinity chromatography as the first step. The sample was fairly impure even after using the ion exchange column as the second step. To remove impurities HIC was done similar to the purification scheme of EcoP1I$^{N\text{-}His}$ (Figure 2.4, F). SEC was used as the final step of purification of all constructs resulting in 99% homogeneous sample suitable for crystallographic studies (Figure 2.4, A, D, E, F).

3.3 DNA binding characteristics

Subsequent to purification of various constructs of EcoP1I detailed in section 3.2, the nuclease dead mutant EcoP1I(E916A)$^{C\text{-}His}$ and EcoP1I$^{N\text{-}His}$ were subjected to binding studies with a variety of DNA substrates including long linear, supercoiled (SC) and short oligomers.

3.3.1 Binding of EcoP1I (E916A)$^{C\text{-}His}$ to long linear substrates

We carried out DNA binding studies of EcoP1I(E916A)$^{C\text{-}His}$ with 891/1045_P1 both in absence (Figure 2.5B) and presence (Figure 2.5C) of ATP in buffer C+. As the enzyme concentration increased from 0 to 500 nM, the EcoP1I-DNA complex migrated slower in the gel indicating binding of enzyme to DNA. It was observed that the complexes migrated much slower in absence of ATP than in its presence. At a concentration of 500 nM EcoP1I and in the absence of ATP the EcoP1I-DNA complex migrated at around 7 kb (Figure 2.5B, Lane 12), while at the same enzyme concentration but in presence of ATP the complex migrated at around 3 kb (Figure 2.5C, Lane 12). Similarly, a 1100 bp substrate having no target site for EcoP1I was chosen as non-specific substrate to study DNA binding properties of EcoP1I (Figure 2.5D,E). EcoP1I bound to non-specific DNA. Analogous to specific substrates, we observed effect of ATP on the mobility of EcoP1I-DNA complexes. Without ATP, the 1.1 kb substrate shifted to maximum 6 kb (Figure

2.5D, Lane 12); whereas with ATP, the highest shift was up to 4 kb (Figure 2.5E, Lane 12).

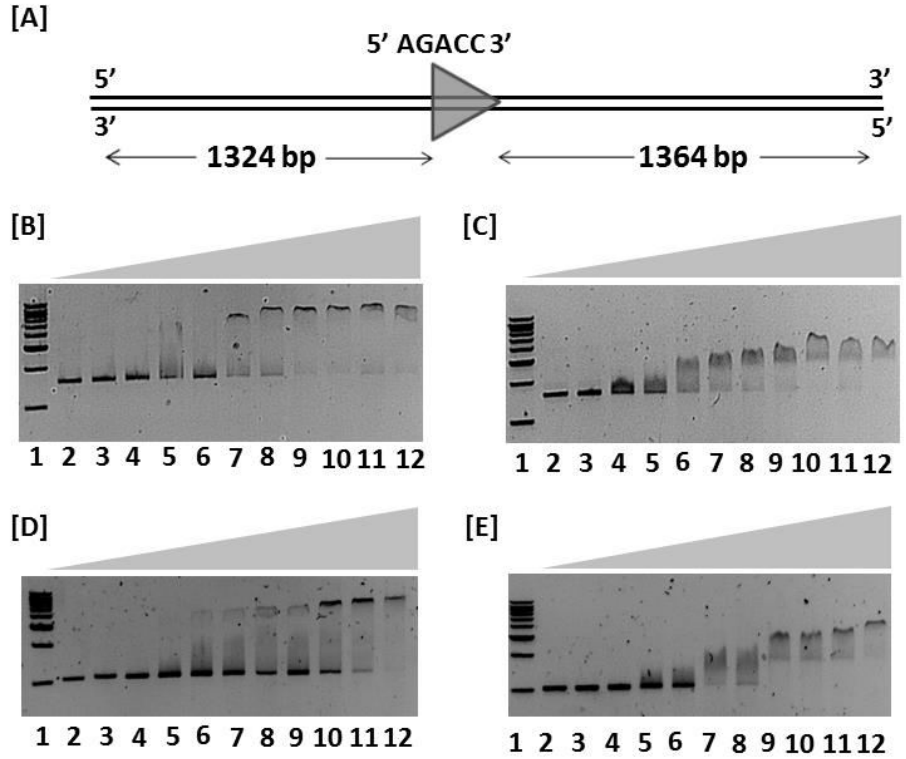


Figure 2.5: **Effect of ATP on DNA binding affinity of EcoP1I(E916A)$^{C\text{-}His}$.** DNA (10 nM) was incubated in buffer C+ with 0, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500 nM of EcoP1I (Lanes 2-12) in absence (panels B,D) or in presence (panels C,E). A] Schematic of specific substrate 891/1045_P1 for EcoP1I binding assay. The target site for EcoP1I (5' AGACC 3') is shown as triangle; B] EMSA with 891/1045_P1 in absence of ATP; C] with 891/1045_P1 in presence of ATP; D] with non-specific DNA substrate in absence of ATP; E] with non-specific DNA substrate in presence of ATP. Lane 1 in B, C, D, E is Genei StepUp™ 1 kb DNA ladder.

To see the effect of cations on the DNA binding affinity of EcoP1I(E916A)$^{C\text{-}His}$, we carried out EMSA in buffer D+ which is potassium based and buffer C+ which is sodium based. The assays were carried out with 891/1045_P1 DNA substrate both, in absence (Figure 2.6A) and presence (Figure 2.6B) of ATP. A similar binding to specific DNA was noticed irrespective of the cation used, i.e. sodium or potassium. At protein concentration above 400 nM, EcoP1I-DNA complex migrated close to 7 kb in absence of ATP (Figure 2.6A, Lanes 5-7 and 12-14), and in presence of ATP the complex migrated close to 3 kb (Figure 2.6B, Lanes 5-7 and 12-14).

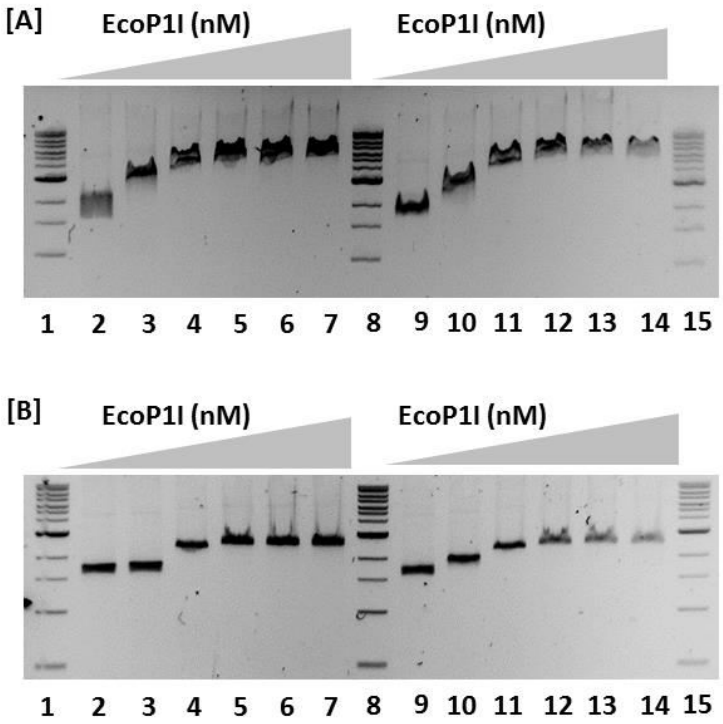


Figure 2.6: Effect of cations on DNA binding affinity of EcoP1I(E916A)$^{C\text{-}His}$. The figure shows EMSA with 891/1045_P1 in absence [A] and presence [B] of ATP for 0, 100, 200, 400, 600 and 800 nM EcoP1I (Lanes 2-7 and 9-14). Reactions contained 10 nM DNA in buffer C+ (Lanes 9-14) or D+ (Lanes 2-7). Lanes 1, 8 and 15 contain 1 kb DNA ladder (NEB).

3.3.2 Binding of EcoP1I (E916A)$^{C\text{-}His}$ to supercoiled (SC) substrates

To see the effect of topology of DNA on the binding of EcoP1I(E916A)$^{C\text{-}His}$, we studied DNA binding properties of EcoP1I to supercoiled and linear DNA substrates. For this study SC pUC18 (Figure 2.7A) and BamHI linearized pUC18 (Figure 2.7B) were used as DNA substrates. DNA binding affinity was judged by examining the "half shift", defined as the protein concentration at which approximately 50% DNA was bound by protein resulting in a shift. Half shift could be seen at around 300 nM EcoP1I for both SC and linear substrates (Figure 2.7, Lane 8 in A and B). This meant that the apparent K_D of the enzyme did not change for either of the two topological forms, indicating binding of EcoP1I to DNA was not influenced by its topology.

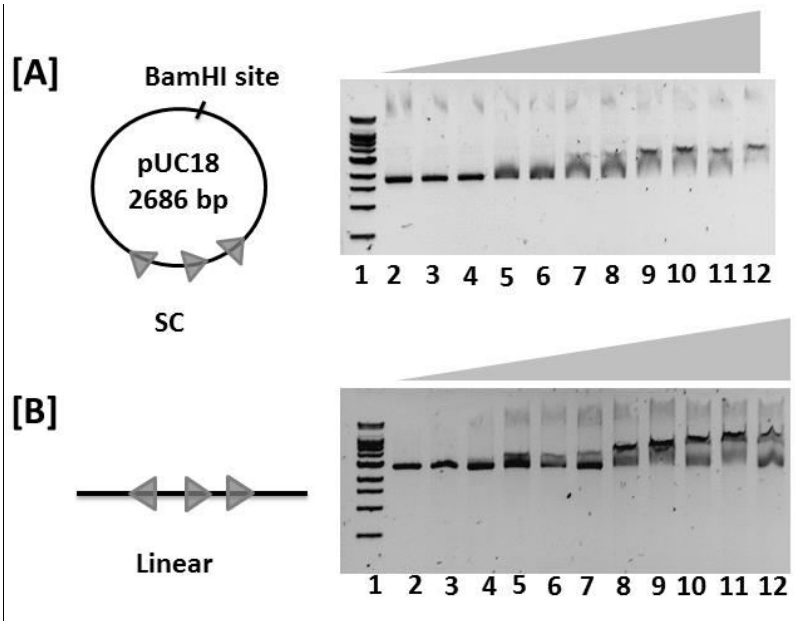


Figure 2.7: DNA binding affinity of EcoP1I with supercoiled DNA and linear DNA substrate. SC (A) or linear (B) DNA (10 nM) was incubated in buffer D+ with 0, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 1000 nM of EcoP1I (Lanes 2-12). Lane 1 contains 1 kb DNA ladder (NEB).

3.3.3 Binding of EcoP1I to short oligomers

The binding of purified EcoP1I(E916A)$^{C\text{-}His}$ to long DNA substrates was followed by examining the enzymes affinity for short oligomers. We shortened the 5' end of 32/32_P1 to as low as 2 bp. The aim was identify short oligomers with high affinity for EcoP1I, making them suitable for co-crystallization studies. EcoP1I bound to all the oligomers with similar affinities (Figure 2.8A-D). As was observed earlier in the binding assays with long linear DNA substrates (Figure 2.5), EcoP1I did not show any difference in binding affinity with specific versus nonspecific oligomer in the concentration range of this EMSA.

To see the effect of cofactor ATP on DNA binding affinity of EcoP1I, we did EMSA of EcoP1I with 32/32_P1 and NS_P1 in the presence of ATP (Figure 2.8F). We did not observe any effect on the binding affinity of EcoP1I with 32/32_P1 versus NS_P1 in presence of ATP. However, the short oligomer containing single target site was found cleaved by EcoP1I in the presence of ATP (Figure 2.8F, Lane 3). This cleaved fragment was detected at 1 μM protein concentration. A similar fragment was not observed at protein concentrations of 3 and 5 μM possibly because EcoP1I bound to the cleaved product as well, and shifted it up (Figure 2.8F, Lanes 4, 5). In contrast, NS_P1 was not

cleaved (Figure 2.8F, Lanes 6-9). In addition to the above experiments, we also checked DNA binding affinity of EcoP1I in presence of S-adenosine methionine (AdoMet) and its structural analogue sinefungin (SF). It was observed that in presence of AdoMet the DNA binding was poorer, while presence of SF did not affect DNA binding affinity (data not shown).

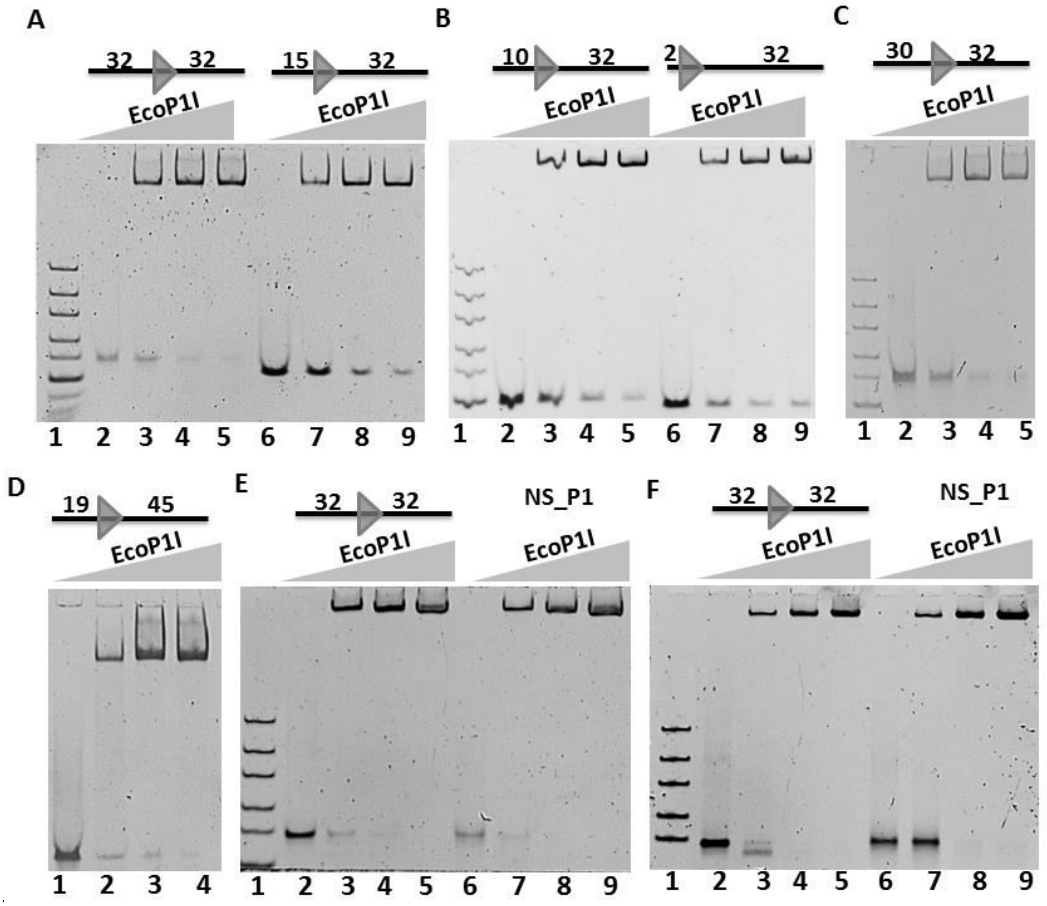


Figure 2.8: EMSA of EcoP1I with short DNA substrates. Protein-DNA complexes were run on 5% native PAGE. EMSA with A] 32/32_P1, Lanes (2-5), 15/32_P1, Lanes (6-9); B] 10/32_P1, Lanes (2-5), 2/32_P1, Lanes (6-9); C] 30/32_P1, Lanes (2-5); D] 19/45_P1; E] 32/32_P1, Lanes (2-5), NS_P1, Lanes (6-9). F] Effect of ATP on binding affinity: 32/32_P1, Lanes (2-5), NS_P1, Lanes (6-9). Lanes (2-5) and (6-9) contain 0,1,3,5 µM EcoP1I in A, B,C,E and F. Lanes (1-4) contain 0,1,3,5 µM EcoP1I in D. Both the cleaved product and the uncleaved DNA can be seen as two lower bands in Lane 3. Lane 1 in A, B, C, E and F contains ultra-low range ladder (NEB).

3.4 ATPase Assay with EcoP1I

The binding assays confirmed that the purified enzyme bound to DNA (Section 3.3). To qualitatively test the ATPase activity of purified EcoP1I, we carried out ATPase assays by malachite green method using SC pUC18 along with short oligomers 15/32_P1 and NS_P1. EcoP1I purified in lab hydrolysed ATP upon incubation with SC pUC18 (Figure 2.9B). With the specific substrate 15/32_P1, EcoP1I rapidly hydrolysed ATP (Figure 2.9C) while with the non-specific DNA NS_P1, the ATPase activity was very low (Figure 2.9C). We observed that the initial rate of ATP hydrolysis was 10 fold lower for nonspecific DNA (2.7 µM Pi/min) than that for specific DNA substrate (20 µM Pi/min).

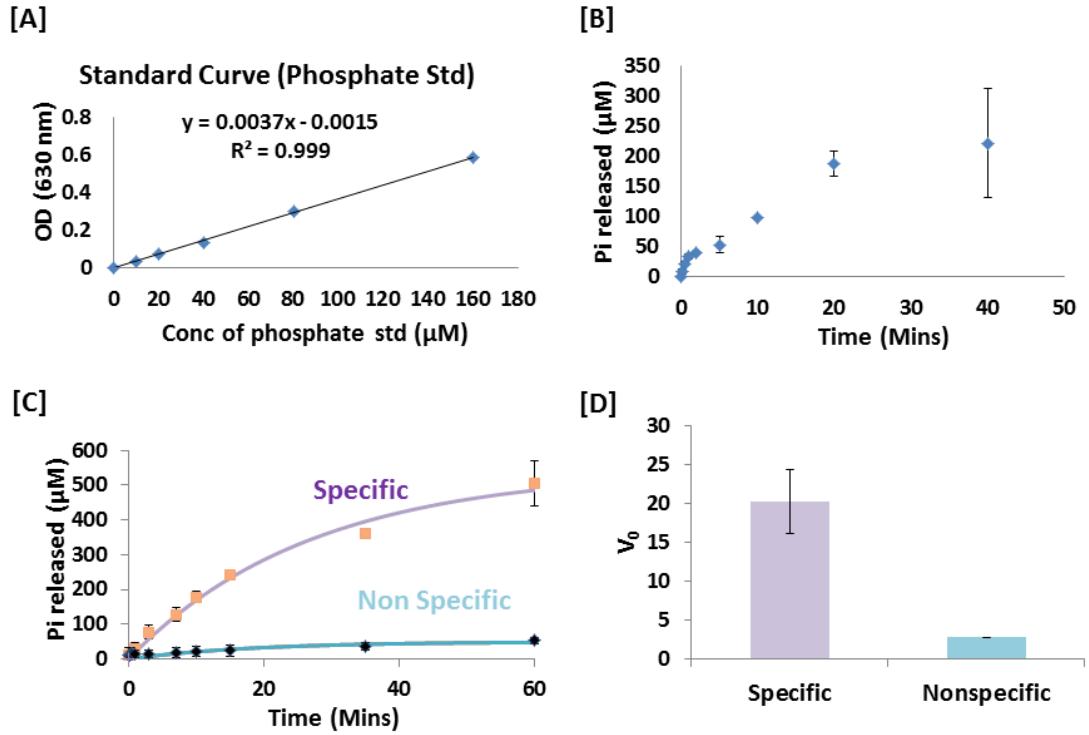


Figure 2.9: ATPase activity of EcoP1I. A] Phosphate standard curve using $(Na)_2HPO_4$; B] ATPase kinetics of EcoP1I on SC pUC18. Reactions contained 10 nM pUC18 and 50 nM EcoP1I in buffer D+; C] ATPase assay with 15/32_P1 (purple) and NS_P1 (cyan). Reactions contained 200 nM DNA and 200 nM EcoP1I in buffer D+; D] Bar graph representing initial velocities (V_0) (µM/min) of ATP hydrolysis for specific (purple bar) and nonspecific (cyan bar) DNA substrates.

ATPase assay was further used to calculate the active fraction of purified enzyme based on the studies carried out by Bianco et al on Type I RM enzyme EcoR124I (20). The active fraction of EcoP1I was determined by titrating EcoP1I relative to a fixed concentration (200 nM) of 15/32_P1. The binding of EcoP1I to this oligomer was confirmed earlier (Figure 2.8B). Furthermore, recognition of the target sequence stimulated the ATPase activity of EcoP1I (Figure 2.9C and D). Consequently, saturation of phosphate release would occur when all the available target sites are bound by EcoP1I. Time course of ATP hydrolysis was plotted for increasing concentrations of EcoP1I (Figure 2.10). Next we plotted the initial rates of the reactions against the corresponding EcoP1I concentration. Figure 2.10F shows that as EcoP1I concentration increased, the rate of reaction (inorganic phosphate released per unit time) steadily increased until 200 nM EcoP1I after which it remained steady at 20 μM/Min. This indicated that ATPase activity saturated at 200 nM EcoP1I when the reactions contained 200 nM of 15/32_P1. Earlier studies have shown that Type III RM enzymes bound to specific target sites have a very low off rate of the enzyme from DNA (21). As the ATP hydrolysis rate saturated at 1:1 stoichiometry of EcoP1I:DNA, the active fraction of EcoP1I in the purified sample is close to 100%.

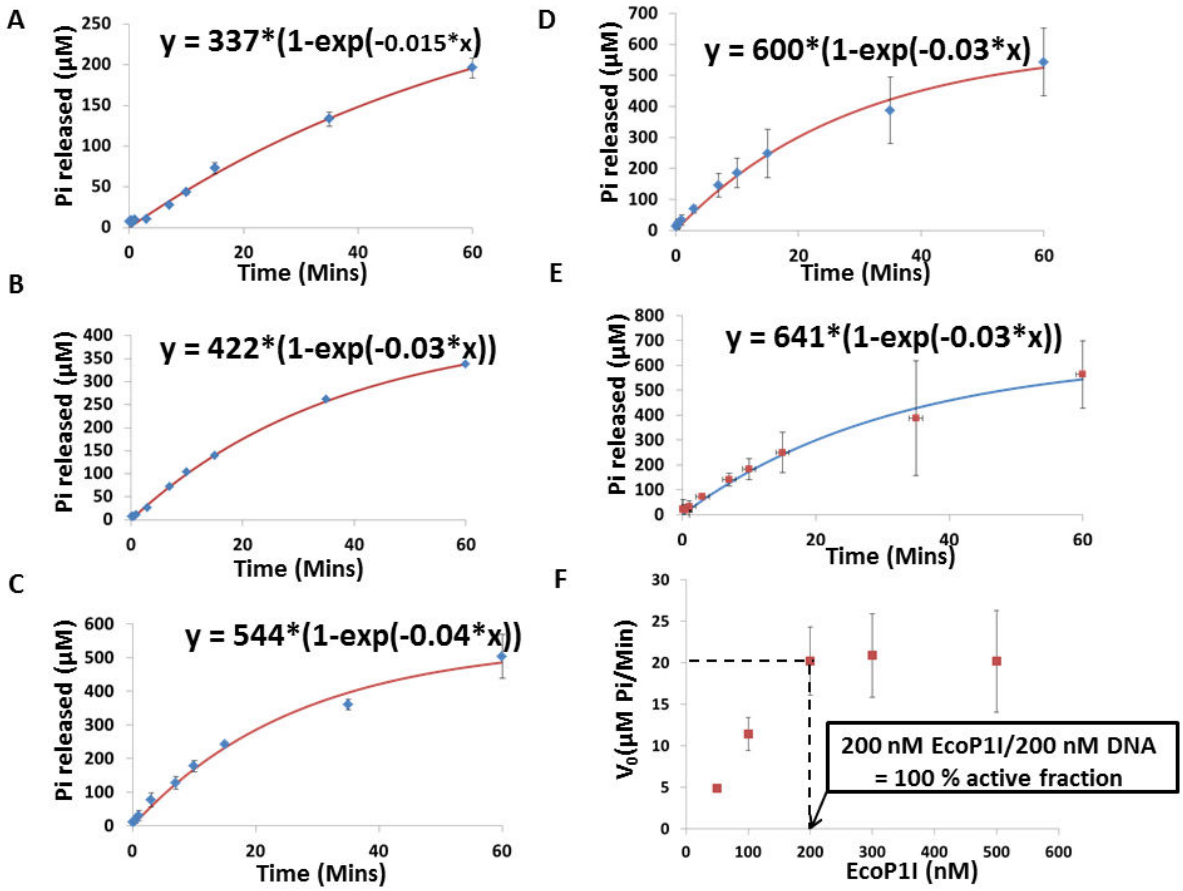


Figure 2.10: ATPase assays of EcoP1I with 15/32_P1. A-E] Time course of ATP hydrolysis. Y axis represents inorganic phosphate concentration (µM) and X axis represents time (minutes). Reactions contained 200 nM 15/32_P1 and 50, 100, 200, 300, 500 nM EcoP1I (A, B, C, D, E), respectively. Reactions were started by addition of ATP and samples (20 µL) were withdrawn at specified time points for analysis. The data was fit (red line) by using raw data (blue squares); the equations of curve are shown on each graph. F] Analysis of time-course shown in A-E to determine the active fraction of EcoP1I. The initial rates of reaction (µM Pi/Min) (y axis) calculated from time-courses of ATP hydrolysis were plotted against respective EcoP1I concentration (nM) (x axis).

3.5 Cleavage assays with long linear and supercoiled substrates having multiple recognition sites

Along with assessing DNA binding and ATPase activity of EcoP1I, we checked the cleavage efficiency of purified EcoP1I with a variety of substrates. As described in section 2.5, we generated circular HtH substrate (pHtH) by silencing one the three target sites in pUC18. On a circular plasmid, HtH is same as TtT substrate (Figure 2.11A). EcoP1I efficiently linearized SC pHtH (Figure 2.11B). The schematic of various possible cleavage products and agarose gel electrophoresis of cleaved products is shown in Figure 2.11A and B respectively. Linearization of pHtH is a result of cleavage

close to one of the two sites resulting in a 2686 bp fragment. Along with this cleavage product, we also observed further processing of pHtH by EcoP1I. The 2686 bp linear fragment represents a linear TtT DNA substrate which further got cleaved by EcoP1I resulting in 2344 bp and 342 bp fragments (Figure 2.11, Lanes 7,8).

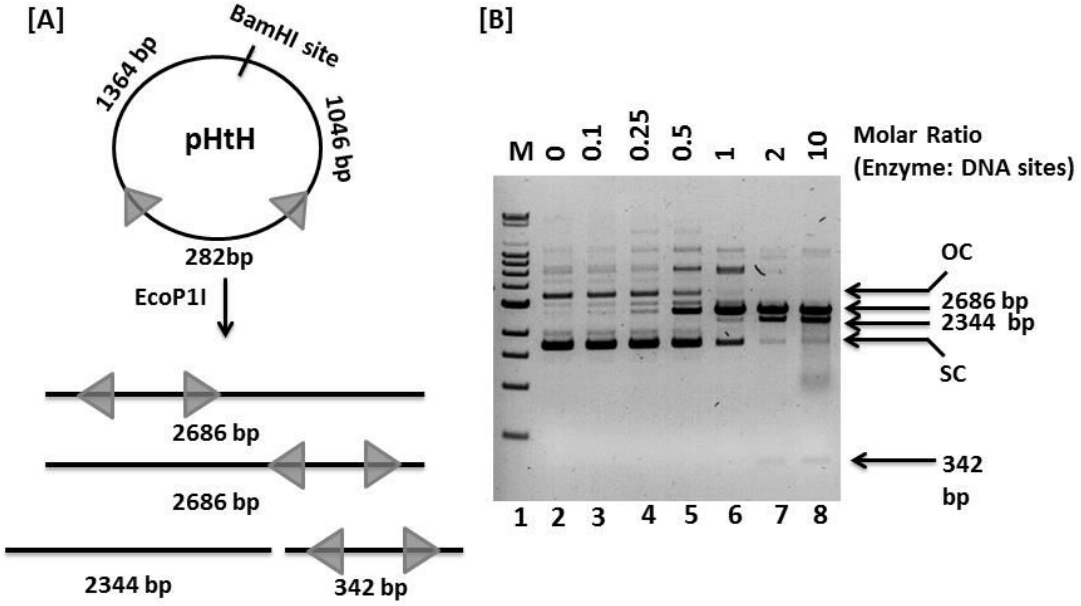


Figure 2.11: Cleavage assay with SC pHtH. A] Possible cleavage products of pHtH after treating with EcoP1I; B] Agarose gel showing cleavage assay with pHtH. Numbers above each well represent molar ratio of Enzyme:DNA sites. Sizes of different cleavage products are indicated on right side. SC: Supercoiled, OC: Open circular. M denotes supermix DNA ladder (Merck Biosciences).

In the next step, HtH, HtT and TtT "linear" substrates were generated from single enzyme restriction digestion of pHtH and pHtT as described in section 2.4. Assuming cleavage downstream of the target sites (1, 2), a schematic of various possible nucleolytic products and the corresponding agarose gels are shown in Figure 2.12. EcoP1I cleaved HtH substrates close to either of the target sites resulting in 2488, 2536, 139 and 187 bp fragments. Simultaneous cleavage close to both target sites resulted in 2354, 139 and 187 bp fragments (Figure 2.12A, D). The 2488, 2536 and 2354 bp fragments co-migrated and could not be resolved. The smaller fragments (139bp and 187 bp) were difficult to visualize due to poor staining by ethidium bromide. However, these fragments could be visualized when the products were analysed using a 10% PAGE gel (data not shown). The TtT substrate was cleaved by EcoP1I close to either of

the target sites resulting in 1339, 1381, 1681, 1039 bp fragments. Simultaneous cleavage close to both target sites resulted in 1339, 1039 and 342 bp fragments (Figure 2.12B, E). The 1339 and 1381 bp fragments co-migrated and were difficult to resolve. The smaller 342 bp fragment was visualized on a 10% PAGE gel (data not shown).

We observed all the possible cleavage products with substrates having inversely oriented sites viz, HtH and TtT. However, cleavage of an HtT substrate was weak and predominantly close to the site shown as orange triangle in Figure 2.12C resulting in co-migrating 1339 and 1381 bp fragments. Cleavage close to target site shown as blue triangle would result in 1626 and 1094 bp fragments; whereas simultaneous cleavage at both target sites would result in 1339, 287 and 1094 bp fragments.

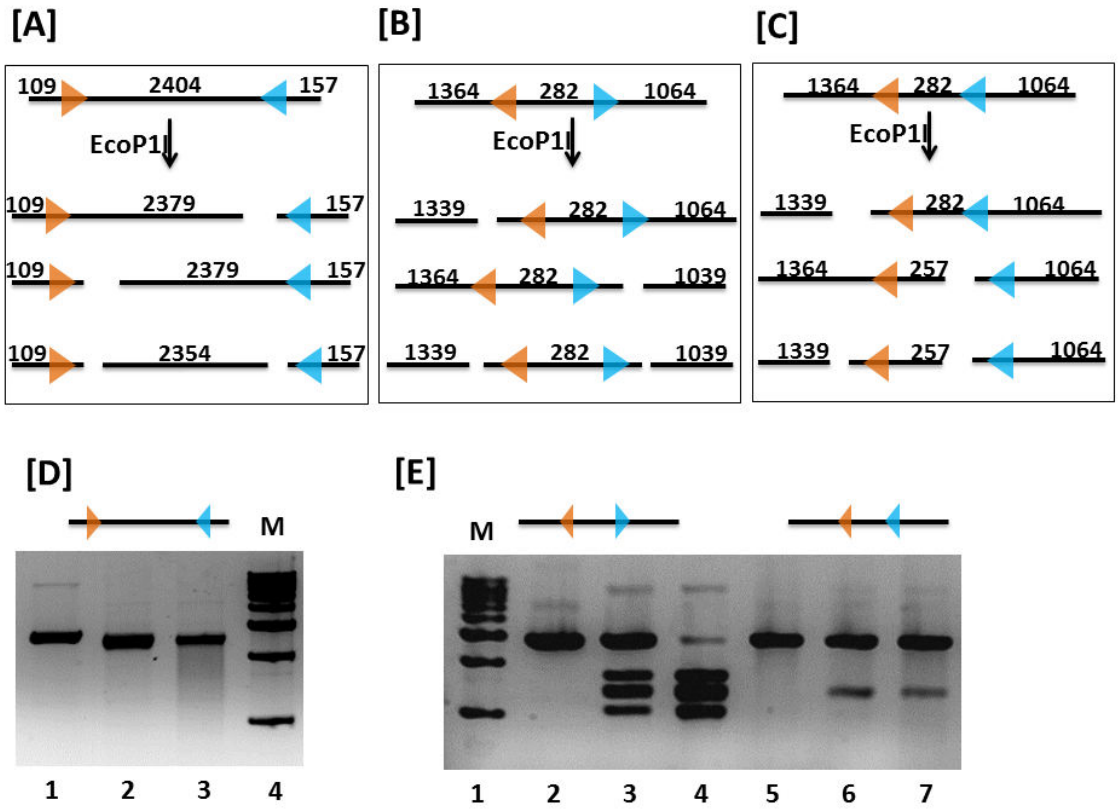


Figure 2.12: Cleavage assay of EcoP1I with various linear substrates. Reactions contained 10 nM DNA incubated with increasing concentration of EcoP1I in buffer D+. Possible cleavage products of HtH (A), TtT (B) and HtT (C) substrates after digestion with EcoP1I are shown as a schematic; D] Cleavage of linear HtH substrate with EcoP1I. Lanes 1-3 depict enzyme:site molar ratio of 0,5,10 respectively. Lane 4 depicts Genei StepUp™ 1 kb DNA ladder. E] Cleavage of linear TtT and HtT substrates with EcoP1I. Lanes 2-4 and 5-7 depict Enzyme:site molar ratio of 0,4,8 respectively. Lane 1 depicts Genei StepUp™ 1 kb DNA ladder.

4. Discussion

Towards purifying a homogeneous sample of EcoP1I in large amounts suitable for crystallization trials, we cloned the appropriate gene/s in high expression vectors. We observed $Mod^{C\text{-}His}$ could be purified by one step purification using affinity chromatography. On the other hand $Res^{C\text{-}His}$ was found predominantly in inclusion bodies, indicating that it is not stable on its own. However, coexpression of Res and Mod yielded soluble EcoP1I. Though we could purify $ATPase^{C\text{-}His}$ by NiNTA affinity chromatography, but SEC studies showed that it was aggregated. This is consistent with previous studies on EcoP15I Res (22, 23). A previous attempt by Wyszomirski et al (22) to purify ATPase domain of EcoP15I by co-expressing with Mod and then removing Mod using high salt concentration resulted in a protein with poor rate of ATP hydrolysis.

In the current structure of EcoP15I, the nuclease domain could not be built because of poor electron density (10). As a strategy for structure determination of the nuclease domain, we tried purifying just the nuclease domain. Proteolysis experiments done with EcoP15I demonstrated the presence of a linker between ATPase and nuclease domains (14). This information combined with extensive sequence analysis of nuclease domain associated with Type III RM enzymes (detailed in Chapter 4) was employed to delineate the boundaries of the ATPase domain, linker and nuclease domain. Based on this information two constructs of the nuclease domain were designed. However, proteins expressed from both the constructs were found insoluble. In summary, we were successful in purifying large quantities of homogenous EcoP1I and Mod. Purification of the enzymes was followed by examination of their activities.

DNA binding properties: The highly purified EcoP1I was able to bind supercoiled as well as linear substrates having EcoP1I target sites. We also compared the binding characteristics of EcoP1I with non-specific long linear and short DNA substrates. Using EMSA, we did not observe any difference in affinities of EcoP1I for specific or non-specific DNA either in presence or absence of ATP. This is surprising since Schwarz et al., noticed using magnetic tweezers assay combined with TIRF that EcoP15I discriminated between specific and non-specific DNA in presence of ATP (6). Furthermore, our results of EcoP1I are inconsistent with the experiment on EcoP15I with radioactively labelled oligomers, which revealed differences in the apparent K_D for

specific and non-specific short DNA substrates (22). It is not clear to us why our assay did not show a difference in binding affinities for specific and non-specific DNA.

Interestingly, in all the binding assays that were carried with long linear DNA, we clearly observed a slower migration of EcoP1I-DNA complexes in the absence than in presence of ATP. This suggests that a larger number of enzymes were bound to the DNA in absence of ATP than in presence. However, this discrimination was not reflected in apparent K_D. Similarly, in binding assays with short oligomers, EcoP1I showed no difference in binding between specific and nonspecific DNA. The assays presented in this study were performed to qualitatively judge the DNA binding ability of purified EcoP1I. Accurate quantitative measurements using sophisticated techniques, such as Isothermal Calorimetry (ITC) or fluorescent anisotropy, can provide correct estimates of binding constants.

As the binding assays were part of preliminary standardization for setting up crystallization trials of EcoP1I-DNA complex, it was also necessary to see the effect of cation on the DNA binding affinity of EcoP1I. Protein-DNA interactions are sensitive to the electrostatic conditions in solution. It has been suggested that cations have a pronounced effect in locating the target site by restriction enzyme EcoRV (24). Also, the Type II restriction enzyme SfiI has been shown to exhibit differential inter-site communication depending on ionic strength of the reaction mixture (25). Previously, Peakman et al studied the effects of sodium and potassium ions on the cleavage activity EcoP1I (12). It was found that in contrast to sodium ions, potassium ions support promiscuous cleavage of DNA substrates. This study focused on effect of monovalent cations on cleavage activity of EcoP1I. To see what effect these cations have on the binding affinity by EcoP1I, we performed EMSA of EcoP1I in sodium and potassium based buffers. Our data demonstrated that replacement of sodium ions with potassium ions had no drastic effect on DNA binding. This information was used later to fine-tune the quality of EcoP1I-DNA crystals (See Chapter 5).

Type III RM enzymes “read” the target sequence to carry out further processes such as activation of ATPase and cleavage of DNA downstream of target site (10). The readout mechanisms are widespread in DNA binding proteins, including base readout (chemical

signatures associated with bases) and shape readout (shape of DNA contour or topology) (26). The geometry of a linear DNA is potentially different from a compact supercoiled DNA. Supercoiling not only compacts the DNA but also brings two distant locations on the DNA closer. Many DNA binding proteins bind preferentially to negatively supercoiled DNA (27). Prokaryotic DNA helicase RecQ which also belongs to SF2 helicases, type IIA topoisomerase IV and mitochondrial transcription factor Tfam have been shown to prefer supercoiled DNA substrates. To see if EcoP1I, showed any preference to topology of DNA; we carried out EMSA using SC pUC18 and linearized pUC18. EcoP1I did not show any preference to supercoiled DNA. The binding assays done with EcoP1I here are not very sensitive to probe the accurate DNA binding affinity of enzyme. Further detailed investigation may be required to conclusively show the difference in binding affinities of EcoP1I towards supercoiled versus linear DNA.

ATP hydrolysis by EcoP1I: Type III RM enzymes bring about mechanical motion on DNA by hydrolysing ATP, though 1000 fold less than closely related Type I RM systems (7, 13, 28, 29). We carried out ATPase assays to check whether the ATPase activity of our enzyme preparation was intact. Further we used ATP hydrolysis kinetics to determine the active fraction of our EcoP1I preparation based on experiments done by Bianco et al (20). For this experiment we used short oligomer (15/32_P1) to ensure that only single molecule of EcoP1I binds he DNA. This assumption is based on the footprint of the enzyme on bound DNA (9), SEC-MALS experiments of EcoP1I and EcoP15I bound to short single-target DNA (30), and the crystal structure of Ecop15I-DNA complex (10). Saturation in ATPase rate is expected when one molecule of EcoP1I productively interacts with one molecule of DNA. The ATP hydrolysis rate saturated at 1:1 stoichiometry of enzyme, indicating approximately 100% active enzyme fraction.

DNA cleavage properties: We carried out cleavage assays with EcoP1I on different substrates. These substrates were designed to have all the possible orientation of target sites. EcoP1I efficiently cleaved substrates with inversely oriented sites i.e. HtH and TtT, however cleavage efficiency was slightly lower with TtT substrate. Interestingly, with HtT oriented sites, the enzyme showed a preference to cleave close to one of the sites over the other. Similar findings are reported by Buchner et al, where EcoP15I was shown to cleave HtT substrates albeit less efficiently with a slight preference to cleave

close to one of the sites (31). To check the binding affinity of EcoP1I with long linear substrates, we used EcoP1I(E916A), a nuclease dead mutant (Figures 2.10 to 2.13). However, when binding assays were done with nuclease active EcoP1I; we observed that the single target DNA substrate was also getting cleaved (Figure 2.8F, Lane 3). It is reported that Type III RM enzymes do not cleave substrates with directly repeated target sites or single target site (3, 13). However, such cleavage events are now being observed (9, 12, 31–33). In pursuit to find how heterotrimeric EcoP1I manage to produce a dsDNA break on a single target substrate, we characterised nucleolytic cleavage of single target site by EcoP1I. These efforts are detailed in Chapter 3.

5. References

1. Hadi,S.M., Baechi,B., Shepherd,J.C.W., Yuan,R., Ineichen,K. and Bickle,T.A. (1979) DNA recognition and cleavage by the EcoPl5 restriction endonuclease. *J. Mol. Biol.*, **134**, 655–666.

2. Baechi,B., Reiser,J. and Pirrota,V. (1979) Methylation and cleavage Sequences of the EcoPl Enzyme. *J. Mol. Biol.*, **128**, 143–163.

3. Meisel,A., Bicklel,T.A., Kruger,D.H. and Schroeder,C. (1992) Type III restriction enzymes need two inversely oriented recognition sites for DNA cleavage. *Nature*, **355**, 467–469.

4. Kruger,D.H., Kupper,D., Meisel,A., Tierlich,M., Reuter,M. and Schroeder,C. (1995) Restriction endonucleases functionally interacting with two DNA sites. *Gene*, **157**, 165.

5. Aelst,K. Van, Tóth,J., Ramanathan,S.P., Schwarz,F.W., Seidel,R. and Szczelkun,M.D. (2010) Type III restriction enzymes cleave DNA by long-range interaction between sites in both head-to-head and tail-to-tail inverted repeat. *PNAS*, **107**, 9123–9128.

6. Schwarz,F.W., Toth,J., van Aelst,K., Cui,G., Clausing,S., Szczelkun,M.D. and Seidel,R. (2013) The helicase-like domains of Type III restriction enzymes trigger long-range diffusion along DNA. *Science.*, **340**, 353–356.

7. Bollins,J. and Szczelkun,M.D. (2015) Re-evaluating the kinetics of ATP hydrolysis during initiation of DNA sliding by Type III restriction enzymes ulia oth. *Nucleic Acids Res.*, **43**, 10870–10881.

8. Ramanathan,S.P., Aelst,K. Van, Sears,A., Peakman,L.J., Diffin,F.M., Szczelkun,M.D. and Seidel,R. (2009) Type III restriction enzymes communicate in 1D without looping between their target sites. *PNAS*, **106**, 1748–1753.

9. Merlind,M., Reich,S., Mo,E., Reuter,M. and Kruger,D.H. (2001) DNA Cleavage by type III restriction-modification enzyme EcoP15I is independent of spacer distance between two head to head oriented recognition sites. *J. Mol. Biol.*, **312**, 687–698.

10. Gupta,Y.K., Chan,S.-H., Xu,S. and Aggarwal,A.K. (2015) Structural basis of asymmetric

DNA methylation and ATP-triggered long-range diffusion by EcoP15I. *Nat. Commun.*, **6**, 1–10.

11. Janscak,P., Sandmeier,U., Szczelkun,M.D. and Bickle,T.A. (2001) Subunit assembly and mode of DNA cleavage of the Type III restriction endonucleases EcoP1I and EcoP15I. *J. Mol. Biol.*, **306**, 417–431.

12. Peakman,L.J., Antognozzi,M., Bickle,T.A., Janscak,P. and Szczelkun,M.D. (2003) S-Adenosyl methionine prevents promiscuous DNA cleavage by the EcoP1I type III restriction enzyme. *J. Mol. Biol.*, **333**, 321–335.

13. Saha,S. and Rao,D.N. (1995) ATP hydrolysis is required for DNA cleavage by EcoPI restriction enzyme. *J. Mol. Biol.*, **247**, 559–567.

14. Wagenführ,K., Pieper,S., Mackeldanz,P., Linscheid,M., Krüger,D.H. and Reuter,M. (2007) Structural domains in the type III restriction endonuclease EcoP15I: characterization by limited proteolysis, mass spectrometry and insertional mutagenesis. *J. Mol. Biol.*, **366**, 93–102.

15. Miroux,B. and Walker,J.E. (1996) Over-production of Proteins inEscherichia coli: Mutant Hosts that Allow Synthesis of some Membrane Proteins and Globular Proteins at High Levels. *J. Mol. Biol.*, **260**, 289–298.

16. Ent,F. Van Den and Lo,J. (2006) RF cloning : A restriction-free method for inserting target genes into plasmids. *J. Biochem. Biophys. methods.*, **67**, 67–74.

17. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.*, **74**, 5463–5467.

18. Baykov,A.A., Evtushenko,O.A. and Avaeva,S.M. (1988) A malachite green procedure for orthophosphate determination and its use in alkaline phosphatase-based enzyme immunoassay. *Anal. Biochem.*, **171**, 266–270.

19. Geladopoulos,T.P., Sotiroudis,T.G. and Evangelopoulos,A.E. (1991) A malachite green colorimetric assay for protein phosphatase activity. *Anal. Biochem.*, **192**, 112–116.

20. Bianco,P.R. and Hurley,E.M. (2005) The type I restriction endonuclease EcoR124I , couples ATP hydrolysis to bidirectional DNA translocation. *J. Mol. Biol.*, **352**, 837–

859.

21. Raghavendra,N.K. and Rao,D.N. (2003) Functional cooperation between exonucleases and endonucleasesÐbasis for the evolution of restriction enzymes. *Nucleic Acids Res.*, **31**, 1888–1896.

22. Wyszomirski,K.H., Curth,U., Mackeldanz,P., Schutkowski,M., Kru,D.H. and Mo,E. (2011) Type III restriction endonuclease EcoP15I is a heterotrimeric complex containing one Res subunit with several DNA-binding regions and ATPase activity. *Nucleic Acids Res.*, 10.1093/nar/gkr1239.

23. Redaschi,N. and Bickle,T.A. (1996) Posttranscriptional regulation of EcoP1I and Eco P15I restriction activity. *J. Mol. Biol.*, **257**, 790–803.

24. Gowers,D.M. and Halford,S.E. (2003) Protein motion from non-specific to specific DNA by three-dimensional routes aided by supercoiling. *EMBO J.*, **22**, 1410–1418.

25. Nobbs,T.J. and Halford,S.E. (1995) DNA cleavage at two recognition sites by the SfiI restriction endonuclease: salt dependence of cis and trans interactions between distant DNA sites. *J. Mol. Biol.*, **252**, 399–411.

26. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.*, **79**, 233–269.

27. Holt,I.J., Neuman,K.C., Litwin,T.R. and Sol,M. (2015) A robust assay to measure DNA topology-dependent protein binding affinity. **43**, 1–10.

28. Reiser,J. and Yuan,R. (1976) Purification and properties of the P15 specific restriction endonuclease from Escherichia coli. *J. Biol. Chem.*, **252**, 451–456.

29. Smith,R.M., Josephsen,J. and Szczelkun,M.D. (2009) The single polypeptide restriction – modification enzyme LlaGI is a self-contained molecular motor that translocates DNA loops. *Nucleic Acids Res.*, **37**, 7219–7230.

30. Butterer,A., Pernstich,C., Smith,R.M., Sobott,F., Szczelkun,M.D. and Tóth,J. (2014) Type III restriction endonucleases are heterotrimeric: comprising one helicase-nuclease subunit and a dimeric methyltransferase that binds only one specific DNA. *Nucleic Acids Res.*, **42**, 5139–50.

31. Möncke-buchner,E., Rothenberg,M., Reich,S., Wagenführ,K., Matsumura,H., Terauchi,R., Krüger,D.H. and Reuter,M. (2009) Functional characterization and modulation of the DNA cleavage efficiency of Type III restriction endonuclease EcoP15I in its interaction with two sites in the DNA target. *J. Mol. Biol.*, **387**, 1309–1319.

32. Peakman,L.J. and Szczelkun,M.D. (2009) S-Adenosyl homocysteine and DNA ends stimulate promiscuous nuclease activities in the Type III restriction endonuclease EcoPI. *Nucleic Acids Res.*, **37**, 3934–3945.

33. Raghavendra,N.K. and Rao,D.N. (2004) Unidirectional translocation from recognition site and a necessary interaction with DNA end for cleavage by Type III restriction enzyme. *Nucleic Acids Res.*, **32**, 5703–5711.

CHAPTER 3

Nucleolytic Cleavage of Single-site Substrates by EcoP1I

Chapter 3

Nucleolytic Cleavage of Single-site Substrates by EcoP1I

1. Introduction

Type III RM enzymes represent a unique class of molecular motors by virtue of their ability to communicate over long distances utilizing 1000 fold less ATP than other classical SF2 helicases (1, 2). Cleavage of dsDNA by Type III RM enzymes requires two sites in an inverted orientation, i.e. either in Head-to-Head (HtH) or Tail-to-Tail (TtT) on the same piece of DNA (3). This is referred to as "site orientation selectivity". Site orientation selectivity acts as a filter for identification of correct geometry of substrate DNA molecule, mainly because the orientation of the target site also sets the orientation of loading enzyme molecule. Followed by identification of target site, the enzyme communicates with another target bound enzyme molecule which may be separated by more than thousands of base pairs (3, 4). A number of models have been proposed to illustrate site orientation selectivity and long range communication associated with Type III RM enzymes viz: translocation, looping and collision (5), end reversal (6), transient looping and collision (7) and 1D bidirectional diffusion (3, 4, 8, 9). These models have been discussed at length in Chapter 1. Recently it has been proposed that Type III RM enzymes EcoP1I and EcoP15I pose an additional role of ATP induced molecular switch (2, 9) similar to that of MutS enzymes belonging to ABC transporters (10) and clamp loader complexes from AAA+ ATPases (11). In presence of ATP, the heterotrimeric Mod_2Res assembly undergoes a conformational change from loading mode to a much distinct sliding mode amenable for 1D bidirectional diffusion on the polynucleotide track (9). Further, an enzyme executing one-dimensional diffusion on the dsDNA cooperates with a target-bound enzyme present on the same piece of DNA, to cleave it 25-27 bp away from the target-bound enzyme (9). The dsDNA break results in a product having 5' overhang of 2-3 nucleotides (12, 13).

Although site orientation selectivity is proposed to be an essential requirement for the working of Type III RM enzymes, while studying the binding affinity of EcoP1I with DNA duplexes bearing single target site, we noticed that these duplexes were getting cleaved in presence of ATP. There have been previous reports observing cleavage of DNA substrates containing just single target site (6, 14–19). Most of these observations were made on long stretches of DNA (>1000 bp). A couple of studies also reported cleavage of

smaller oligomers (50-70 bp) containing single site by EcoP15I and EcoP1I (14, 19). The mechanism behind these cleavage events, especially those involving short oligomers, could not be addressed by previous models, as they required cooperation between two enzymes in *cis* . The fact that the footprint of EcoP15I on a bound DNA containing single target site was ~36 bp suggested that only one enzyme bound to such short oligomers (14). This was corroborated by Size Exclusion Chromatography combined with Multi-angle Light Scattering (SEC-MALS) studies, which showed that only one molecule of either EcoP1I or EcoP15I bound to a 50 bp long DNA (19). Hence, it is difficult to imagine more than one enzyme molecule assembling on a short oligomer to cooperate in *cis* and catalyzing nucleolytic cleavage. Consequently, a pertinent question regarding the mechanism of single-site cleavage is whether the cleavage is a *cis* or a *trans* event.

It is possible that cleavage of single-site substrate is catalysed by a single enzyme, which nicks both the strands. However, in such a case the nuclease domain of the enzyme would have to be conformationally flexible to bind and nick the second strand. Hence, we feel that the above mentioned models of nucleolytic cleavage fail to explain single-site cleavage. Other possible causes for single-site cleavage include reaction buffer composition (16, 18), enzyme promiscuity (16), or a free enzyme or DNA-bound enzyme cooperating with a target-bound enzyme in *trans* (19). Despite the suggestions, a conclusive evidence for the cause of the non-canonical substrate remained elusive.

To better understand this phenomenon, we characterized the unusual cleavage activity and compared its characteristics with the canonical nucleolytic activity. Furthermore, to find out whether this cleavage event was a *cis* or *trans* event we carried out heparin trap assays and developed a new assay exploiting the cooperation between the heterologous enzymes EcoP1I and EcoP15I to obtain conclusive evidence.

This chapter describes the methods used, results obtained during characterisation of single-site cleavage, and presents a new mode of cleavage by Type III R-M enzymes.

2. Materials and Methods

2.1 DNA substrates

Oligomers were purchased from Integrated DNA Technologies, USA, and Sigma-Aldrich, USA. DNA duplexes were made by annealing the single-strands and further purified using an 8 ml MonoQ™ 10/100 GL column. The resulting duplexes were washed thoroughly with MilliQ-purified water to remove any salt and concentrated using Vivaspin®2 (GE Healthcare) concentrator (MWCO 3kDa). The concentrated oligomer was stored in MilliQ-purified water at -80°C until further use. A list of dsDNA used for the various biochemical assays described below is given in Table 3.1. The T (top strand) and B (bottom strand) pairs were annealed to get a duplex DNA. The convention to label the DNA substrate is as described in section 2.4, Chapter 2.

Table 3.1: Oligomers used for biochemical assays with EcoP1I and EcoP15I*

Name	Sequence (5' -> 3')
32/32_P1_T	CATGACTGAGTCAATCGGATCGTAGACGTACTAGACCTATCCTGTATGCTACGTATTCGTATCGTGAGC
32/32_P1_B	GCTCACGATACGAATACGTAGCATACAGGATAGGTCTAGTACGTCTACGATCCGATTGACTCAGTCATG
30/32_P1_T	TGACTGAGTCAATCGGATCGTAGACGTACTAGACCTATCCTGTATGCTACGTATTCGTATCGTGAGC
30/32_P1_B	GCTCACGATACGAATACGTAGCATACAGGATAGGTCTAGTACGTCTACGATCCGATTGACTCAGTCA
15/32_P1_T	GATCGTAGACGTACTAGACCTATCCTGTATGCTACGTATTCGTATCGTGAGC
15/32_P1_B	GCTCACGATACGAATACGTAGCATACAGGATAGGTCTAGTACGTCTACGATC
10/32_P1_T	TAGACGTACTAGACCTATCCTGTATGCTACGTATTCGTATCGTGAGC
10/32_P1_B	GCTCACGATACGAATACGTAGCATACAGGATAGGTCTAGTACGTCTA
5/32_P1_T	GTACTAGACCTATCCTGTATGCTACGTATTCGTATCGTGAGC
5/32_P1_B	GCTCACGATACGAATACGTAGCATACAGGATAGGTCTAGTAC
2/32_P1_T	CTAGACCTATCCTGTATGCTACGTATTCGTATCGTGAGC
2/32_P1_B	GCTCACGATACGAATACGTAGCATACAGGATAGGTCTAG
5/25_P1_T	GTACTAGACCTATCCTGTATGCTACGTATTCGTAT

5/25_P1_B	ATACGAATACGTAGCATACAGGATAGGTCTAGTAC
5/24_P15_T	GTACTCAGCAGTATCCTGTATGCTACGTATTCGTAT
5/24_P15_B	ATACGAATACGTAGCATACAGGATACTGCTGAGTAC
3/17_P15_T	ACTCAGCAGTATCCTGTATGCTACGT
3/17_P15_B	ACGTAGCATACAGGATACTGCTGAGT
3/16_P15_T	ACTCAGCAGTATCCTGTATGCTACG
3/16_P15_B	CGTAGCATACAGGATACTGCTGAGT
3/11_P15_T	ATACAGCAGTAGACTATGAT
3/11_P15_B	ATCATAGTCTACTGCTGTAT
NS_P1_T	GGCTCACGCTACGACTACGTAGCATACAGGTGACCGTCTAGTACGTCTACGTTCCGAGTGACTCAGTCA
NS_P1_B	TGACTGAGTCACTCGGAACGTAGACGTACTAGACGGTCACCTGTATGCTACGTAGTCGTAGCGTGAGCC
NS_P15_T	ATACGAATACGTAGCATACAGCCTAGATTAGGTAC
NS_P15_B	GTACCTAATCTAGGCTGTATGCTACGTATTCGTAT

*Target sites for EcoP1I/P15I are underlined.

Circular DNA substrate containing single site (pOne) was generated from pHtH by site directed mutagenesis. Primer designing and mutagenesis experiment were done exactly as described in section 2.1.3 (Chapter 2). The primers used for mutagenesis are listed in Table 3.2.

Table 3.2: Primers used for generating pOne from pHtH

Name	Sequence (5' -> 3')
PUC18_LastSil_F	GATCCTTTGATCTTTTCTACGGTCGGGGACGCTCAGTGGAACGAAAACTC
PUC18_LastSil_R	GAGTTTTCGTTCCACTGAGCGTCCCCGACCGTAGAAAAGATCAAAGGATC

2.2 Protein purification

EcoP1I, EcoP1I (E916A) and Mod were purified as described in section 2.3, Chapter 2. EcoP15I was purified in the laboratory by Ishtiyaq Ahmed Khan by a protocol similar to that of EcoP1I.

2.3 Electrophoeretic mobility shift assays (EMSA)

The protein and DNA mix was incubated for 40 minutes at 25° C in buffer D+ in the presence of 20 μM sinefungin (SF) (Sigma). Subsequent to incubation, a native stop buffer (10 mM Tris, pH 8, 60 mM EDTA, 60% glycerol, 0.03% bromophenol blue, 0.03% xylene cyanol) was added and immediately loaded on a 5% native PAGE gel (5% acrylamide:bis-acrylamide(29:1), 1X TBE). The gels were run at 4° C in 1X TBE, stained with a solution containing ethidium bromide and scanned using Typhoon TRIO+ variable mode imager (GE Healthcare).

2.4 Cleavage assays

Cleavage assays were carried out at 25° C in buffer D+ (50 mM Tris pH 8, 50 mM KCl, 10 mM $MgCl_2$, 1 mM DTT, 100 μg/ml BSA). The DNA and protein were incubated in the presence of 20 μM SF. After 40 minutes, 1 mM ATP (Sigma) was added and incubated further for 15 minutes. The reactions were stopped by adding 0.5 volumes of a stop buffer (10 mM Tris pH8, 60 mM EDTA, 60% glycerol, 0.025% SDS, 0.03% bromophenol blue, 0.03% xylene cyanol) and loaded on 15% native PAGE gel (15% Acrylamide:BisAcrylamide (29:1), 1X TBE). For denaturing urea-formamide gel runs, the reactions were stopped by adding 1 volume of formamide gel loading dye (95% formamide, 0.5 mM EDTA, 0.025% SDS, 0.03% bromophenol blue, 0.03% xylene cyanol). The samples were heated at 99° C for 10 minutes, spun briefly and loaded on a 15% urea-formamide PAGE gel [15% acrylamide:bis-acrylamide (29:1), 8 M urea, 20% formamide, 1X TBE]. The gels were stained with a solution containing ethidium bromide and scanned using Typhoon TRIO+ variable mode imager. The DNA bands were quantified by Image Quant v7 (GE Healthcare).

2.5 Methylation assays

The methylation assays were carried out in buffer D+. DNA and protein (Mod or EcoP1I) were incubated in the presence of 100 μM AdoMet (Sigma-Aldrich) for 60 minutes. The

reactions were performed in duplicates. In one of the sets, Mod and AdoMet were removed by cleaning the DNA using PCR purification kit (Qiagen) before subsequent processing.

2.6 ATPase assays

The reactions for the ATPase assay were done similar to cleavage assay in buffer D+. The reactions were stopped at fixed time points by addition of 100 mM EDTA. ATPase activity was checked using malachite green (20, 21). Malachite green solution was prepared by mixing 0.044 g malachite green to 36 mL 3 N H_2SO_4. For the detection of inorganic phosphate, the dye solution was prepared freshly by mixing 800 µL of malachite green solution, 200 µL 7.5% ammonium molybdate and 16 µL of 11% Tween20. For every 20 µL reaction, 50 µL malachite green dye was added. The absorbance at 630 nm was recorded after 15 minutes using Varioskan (Thermo Fischer Scientific). The inorganic phosphate released was calculated from a standard phosphate curve using Na_2HPO_4 plotted for every assay. For data analysis, the data was fit using CurveExpert Professional (Version 2.4.0) and initial rates of reaction were calculated by determining the slope of initial linear portion of time course.

2.7 Heparin trap assays

Heparin trap assays were carried out at 25° C in buffer D+. The dsDNA 15/32_P1 and EcoP1I were incubated in the presence of 20 µM SF for 40 minutes. 0.6% w/v heparin (Sigma) was then added and reaction further incubated for 5 minutes. Finally 1 mM ATP was added. To monitor time dependent cleavage, reaction was stopped at different time points with 0.5 volumes of stop buffer and loaded on 15% native PAGE gel. The gels were stained with a solution containing ethidium bromide and scanned using Typhoon TRIO+ variable mode imager. The DNA bands were quantified by Image Quant V7. For data analysis, the data was fit using CurveExpert Professional (Version 2.4.0) and initial rates of reaction were calculated by determining the slope of initial linear portion of time course.

2.8 Heterologous cooperation assays

Cooperation assays were carried out at 25°C in buffer D+. The dsDNA 15/32_P1 and EcoP1I or EcoP1I (E916A) were incubated in the presence of 20 μM SF for 40 minutes. Similarly, EcoP15I and its target containing dsDNA of varying lengths were incubated separately for 40 minutes. These two complexes were then mixed together and 1 mM ATP was added. After 15 minutes, 1 volume of formamide gel loading dye was added. The samples were heated at 99°C for 10 minutes, spun briefly and loaded on a 15% urea-formamide PAGE gel. The gels were stained with a solution containing ethidium bromide and scanned using Typhoon TRIO+ variable mode imager.

3. Results

3.1 Preliminary characterization of single-site cleavage

In order to better understand the cause of cleavage of single-site substrates by EcoP1I, a prototypical Type III enzyme, we compared its DNA cleavage features with those for canonical DNA substrates. As described in Chapter 1, the canonical substrates for Type III enzymes have a pair of inversely oriented sites (HtH or TtT). Type III RM enzymes cleave these substrates 25-27 bp downstream of the target site utilizing ATP. The resulting cleaved piece of DNA has a 5' overhang of 2 nt. The single-site cleavage by EcoP1I was investigated for specificity of target site, nucleotide dependence, locus of dsDNA break, and presence of overhang in the cleaved product.

3.1.1 Effect of length of DNA

We generated three dsDNA of varying lengths, each containing one target site of EcoP1I (Figure 3.1). Cleavage of 5/32_P1 by EcoP1I resulted in a ~36 bp and ~6 bp fragment. The 6 bp fragment could not be visualized on staining with ethidium bromide (Figure 3.1A). For 5/32_P1, all of the template got cleaved at Enzyme:Site (E:S) ration of as low as 0.66 (Figure 3.1A, Lane 4). Cleavage of 152/192_P1 by EcoP1I resulted in a ~167 bp and ~183 bp fragments (Figure 3.4B). The oligomers were almost fully cleaved at an E:S ratio of 3.2 (Figure 3.1B, Lane 6). Cleavage of a longer DNA substrate 1324/1364_P1 was less efficient than the other two oligomers for same value of E:S, and resulted in co-migrating ~1339 bp and ~1354 bp fragments (Figure 3.1C). Cleavage reaction did not go to completion even at E:S ratio of 16 (Figure 3.1C, Lane 7). Our results suggested that cleavage of short oligomers was more efficient than longer ones.

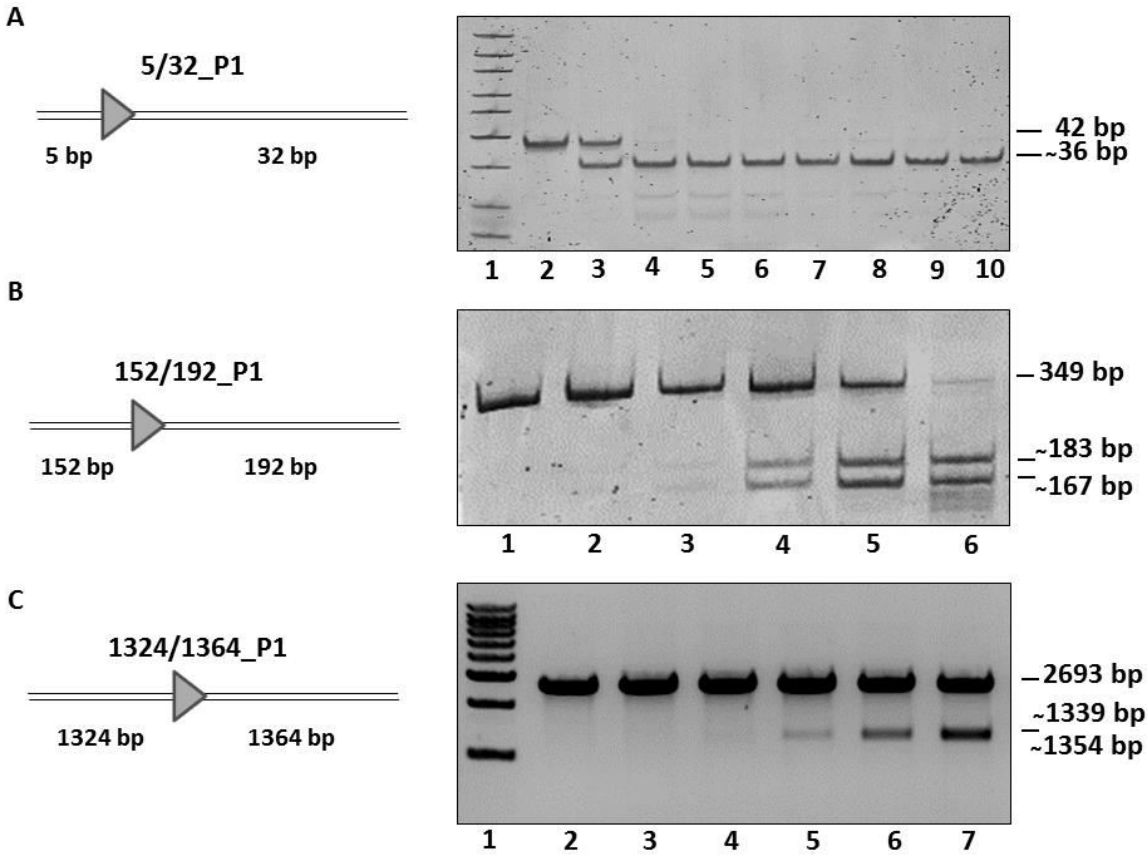


Figure 3.1: Single-site dsDNA got cleaved by EcoP1I irrespective of length of target DNA. A] Cleavage assay with 300 nM 5/32_P1. Lanes 2-10 depict an E:S ratio of 0, 0.33, 0.66, 1, 1.33, 1.66, 2, 2.66, and 3.33. Lane 1 contains GeneRuler Ultra Low Range DNA Ladder (NEB); B] Cleavage assay with 50 nM 152/192_P1. Lanes 1-6 depict an E:S ratio of 0, 0.2, 0.4, 0.8, 1.6, 3.2; C] Cleavage assay with 10 nM 1324/1364_P1. Lanes 2-7 depict an E:S ratio of 0, 1, 2, 4, 8, and 16. Lane 1 contains Genei StepUp™ 1 kb DNA ladder.

3.1.2 Effect of buffer conditions

We performed cleavage assays with 32/32_P1 to see the requirement of ATP binding versus ATP hydrolysis. This was done by providing non-hydrolysable analogues of ATP instead of ATP. Cleavage was noticed only in presence of ATP (Figure 3.2A, Lane 7), and was not observed in absence of ATP (Figure 3.2A, Lane 2) or with ATP analogue AMP-PNP (Lane 4), or with transition state ATP mimic ADP-vanadate (Lane 5) or with ADP (Lane 6). Additionally, DNA cleavage occurred only when the target site was present, non-specific DNA was not cleaved by EcoP1I in presence of ATP (Figure 3.2A, Lane 3).

Next, we checked the effect of monovalent cations on the cleavage of 32/32_P1 in presence of either S Adenosine methionine (AdoMet) or its structural analogue sinefungin (SF) (Figure 3.2B). In presence of KCl, apart from the primary cleavage

product (43 bp), a lot of non-specific cleavage events (promiscuity) could be observed (Figure 3.5B, Lane 2). After incubation with AdoMet, this promiscuity did not go down; however, a decrease in overall cleavage was observed (Figure 3.2, Lane 3). We believe that this is due to methylation of the oligomer by AdoMet. With SF, the cleavage was efficient and specific (Figure 3.2B, Lane 4). With NaCl, overall cleavage efficiency seemed to be lesser, which enhanced on addition of SF. Irrespective of the cation, we noticed that cleavage specificity and efficiency improved on addition of SF (Figure 3.2B, Lanes 5-7).

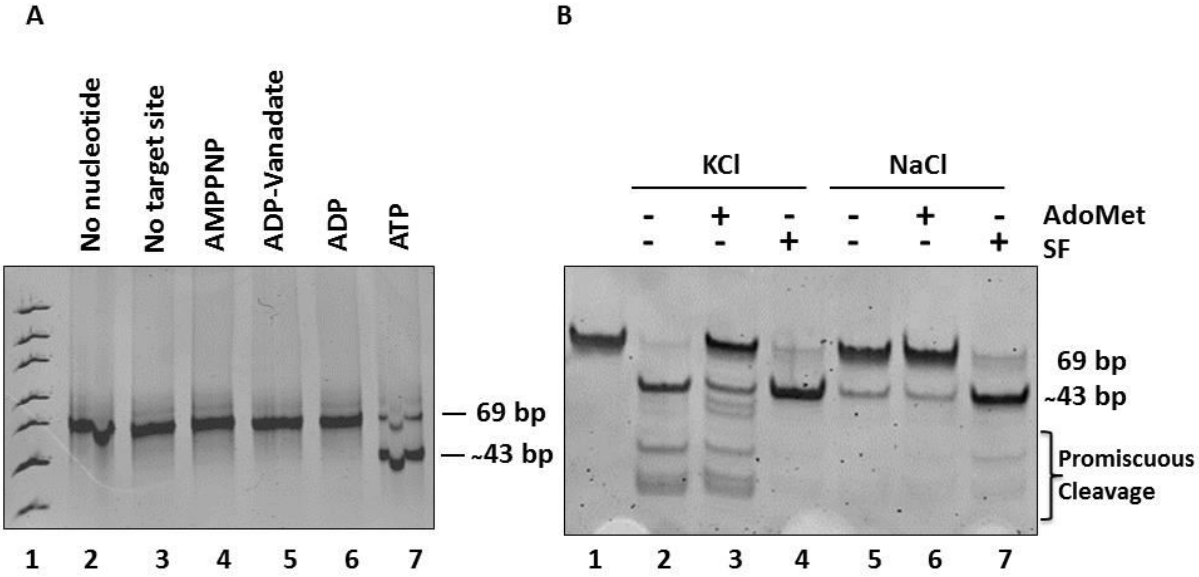


Figure 3.2: Effect of buffer conditions on single-site cleavage by EcoP1I. Reactions contained 300 nM DNA and 200 nM EcoP1I in buffer D+. A] Cleavage assay with various analogs of ATP. 1) GeneRuler Ultra Low Range DNA Ladder (NEB), 2) Without nucleotide, 3) NS_P1+ ATP, 4) 32/32_P1 + AMP-PNP, 5) 32/32_P1 + ADP-Vanadate, 6) 32/32_P1 + ADP, 7) 32/32_P1 + ATP. B] Effect of monovalent cations and AdoMet and SF. 1) Template DNA without EcoP1I, 2) – AdoMet, -SF,+KCl, 3) + AdoMet, -SF, +KCl, 4) – AdoMet, +SF, +KCl, 5) – AdoMet, -SF,+NaCl, 3) + AdoMet, -SF, +NaCl, 4) – AdoMet, +SF, +NaCl. On the right side, sizes of DNA fragments are given. Promiscuous cleavage is shown encompassing the curved bracket on right side.

3.1.3 Methylation by EcoP1I silences the endonuclease activity

We first methylated 32/32_P1 by incubating with Mod and AdoMet. Subsequently ATP and EcoP1I were supplied to initiate cleavage. EcoP1I did not cleave DNA substrates that were incubated with Mod + AdoMet *a priori* (Figure 3.3, Lanes 2, 3, 6, and 7) compared to control reaction where the substrate was not previously incubated with Mod + AdoMet (Figure 3.3, Lane 8). The silencing of the endonuclease activity suggested

that the target site was modified in presence of Mod + AdoMet, and demonstrated that the single-site cleavage was sensitive to the methylation status of the target site.

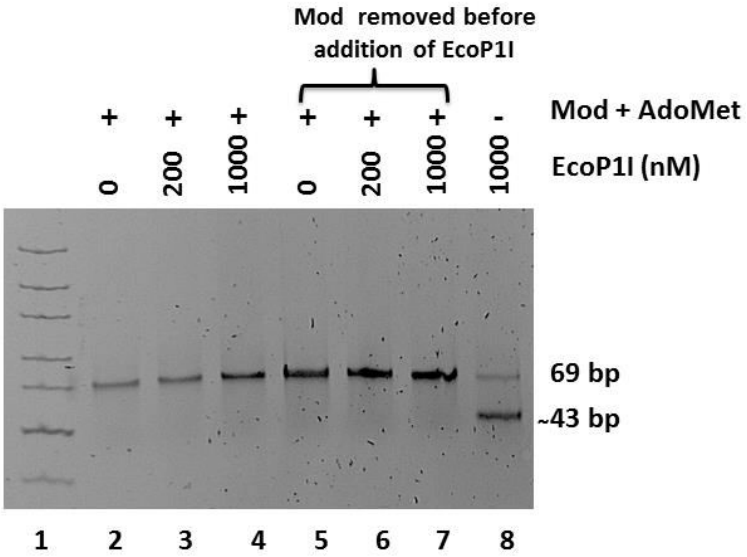


Figure 3.3: Methylation by EcoP1I silences the endonuclease activity. Reactions contained 300 nM 32/32_P1 and 300 nM EcoP1I in buffer D+. Numbers above each well depict EcoP1I concentration in nM.1) GeneRuler Ultra Low Range Ladder (NEB), Lanes 2-4) Mod + AdoMet still present, Lanes 5-7) Mod + AdoMet removed by PCR purification kit before addition of EcoP1I, 8) Control reaction with only EcoP1I.

3.1.4 Effect of methylation on ATPase activity of EcoP1I with single-site DNA

As described in Chapter 2, section 3.4, a single-site substrate 15/32_P1 stimulated the ATPase activity of EcoP1I. Further we carried out extensive ATPase assays with EcoP1I to see the extent of ATP hydrolysis in following conditions:

a) Only 15/32_P1 (representing specific DNA substrate)
b) 15/32_P1 incubated with Mod + AdoMet prior to addition of EcoP1I and ATP (representing methylated DNA substrate)
c) AdoMet and ATP added together (representing a competition between methylation and ATP hydrolysis)
d) Only NS_P1 (representing non-specific DNA)

With specific substrate, EcoP1I rapidly hydrolysed ATP with a V_0 of 20.2 µM Pi/min. With non-specific DNA, ATPase activity was almost 10 fold lower (V_0 of 2.7 µM Pi/min). It was observed earlier that the substrate incubated with Mod + AdoMet before addition of EcoP1I and ATP did not get cleaved (Figure 3.3). To see if methylation hampered the

ATPase activity of EcoP1I; we performed ATPase assays when AdoMet was either added with or before ATP. In each of these cases, the ATPase activity was lower than that with unmodified DNA. The initial velocities for both these conditions were comparable to that for non-specific DNA. This indicated that a DNA with unmodified target site is required for stimulating the ATPase activity. These results are summarized in Figure 3.4.

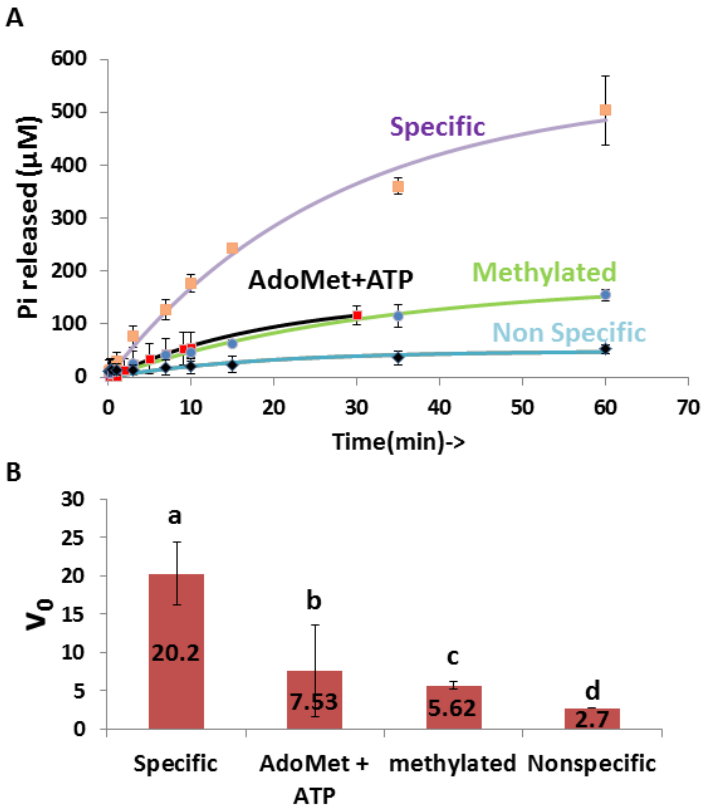


Figure 3.4: ATPase activity of EcoP1I on single-site DNA substrate. Reactions contained 200 nM DNA and 200 nM EcoP1I in buffer D+. A] ATPase assay with 15/32_P1 (purple), NS_P1 (cyan), methylated 15/32_P1 (green) and with 15/32 when AdoMet and ATP were added together (black); B] Bar graph representing initial velocities (V_0) (µM/min) of ATP hydrolysis: a) 15/32_P1, b) 15/32 when AdoMet and ATP were added together, c) methylated 15/32_P1, d) NS_P1. Numbers on bar graph indicate mean values of V_0 after analyzing three experimental replicates.

3.1.5 Location of dsDNA break

To understand the position of dsDNA break on a single-site DNA substrate, we systematically reduced the length of DNA upstream of the target site. Analysis of these different oligomers on a polyacrylamide gel showed that reducing the length upstream of target site did not affect the cleavage of single-site substrates. Cleavage of 30/32_P1 (total length = 67 bp) resulted in ~60 bp fragment (Figure 3.5A, Lanes 6, 7). Cleavage of 15/32_P1 (total length = 52bp) resulted in ~45 bp fragment (Figure 3.5A, Lanes 4, 5). This was also observed for 5/32_P1 (total length = 42bp) which had just 5 bp upstream of the target site. Cleavage of this oligomer resulted in a ~35 bp fragment (Figure 3.5A, Lanes 2, 3). Upon variation of length downstream of the target site, we observed that an oligomer with downstream length < 25 bp was refractory to cleavage (Figure 3.5B). However, we could not discriminate between nicking and dsDNA break of this oligomer on a 15% native PAGE gel. The data conclusively showed that cleavage was taking place downstream of the target site. To exactly map the cleavage position, we sequenced bottom strand of the cleaved product of 2.6 kb DNA substrate (1324/1364_P1) with single target site (Figure 3.5B). The cleavage mapped to 26 nt downstream of the target site for this substrate.

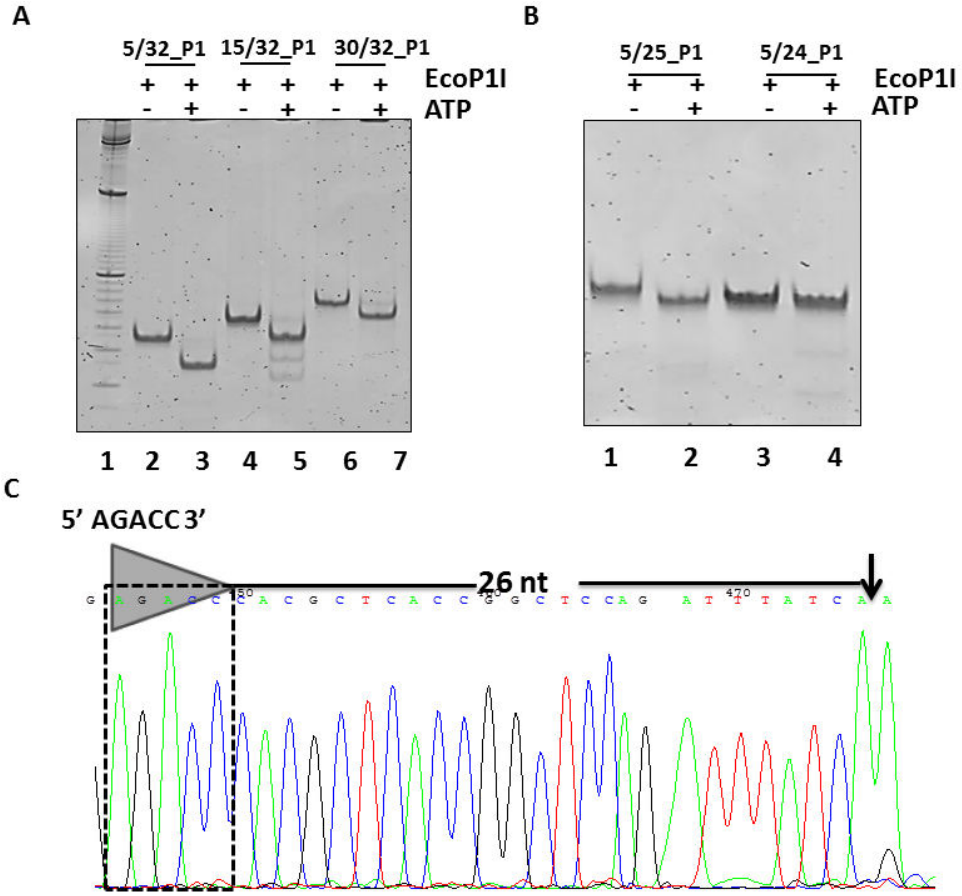


Figure 3.5: Position of dsDNA break. Reactions contained 200 nM DNA and 500 nM EcoP1I in buffer D+. A] Cleavage assay with various single-site oligomers with variation in upstream length. 1) Marker, 2)5/32_P1, +EcoP1I, -ATP, 3) 5/32_P1, +EcoP1I, +ATP, 4) 15/32_P1, +EcoP1I, -ATP, 5) 15/32_P1, +EcoP1I, +ATP, 6) 30/32_P1, +EcoP1I, -ATP, 7) 30/32_P1, +EcoP1I, +ATP; B] Cleavage assay with various single-site oligomers with variation in downstream length. 1) 5/25_P1, +EcoP1I, -ATP, 2) 5/25_P1, +EcoP1I, +ATP, 3) 5/24_P1, +EcoP1I, -ATP, 4) 5/24_P1, +EcoP1I, +ATP; C] Sequencing of bottom strand of cleaved product of 1324/1364_P1. The target sequence of EcoP1I (5' AGACC 3') is highlighted by serrated box and triangle. The black arrow indicates the position of cleavage.

3.1.6 Presence of overhang in the cleaved product

To check if the DNA fragment resulting from cleavage of a DNA with a single target site had an overhang, we denatured the cleaved product. The resulting single strands when run on a denaturing PAGE could give an indication of difference, if any, in their lengths. On the denaturing PAGE, the cleaved product separated into two single strands of DNA with a length difference of approximately 2 nt (Figure 3.6). This observation indicates that the nucleolytic activity of EcoP1I results in an overhang of 2 nt.

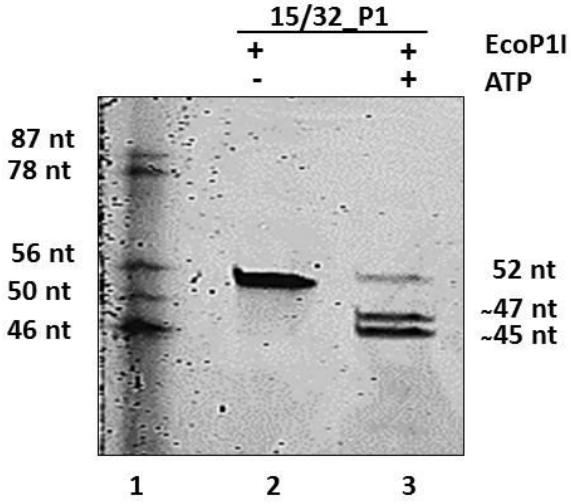


Figure 3.6: The cleaved product has an overhang of ~ 2 nt. 1) Marker, 2) +EcoP1I, -ATP, 3) +EcoP1I, +ATP. The numbers on left and right sides of gel indicate sizes of various DNA fragments in nucleotides.

3.2 Cleavage of DNA with single target site: cis or trans?

After establishing that the properties of single-site cleavage are similar to canonical cleavage, we proceeded to get insights into the mechanism underlying this endonucleolytic activity. There are three possible modes of cleavage of single-site DNA substrates. A single target bound enzyme molecule may execute a dsDNA break in *cis* (Figure 3.7C). An alternate model to cleavage in *cis* is that of cleavage resulting from an enzyme in *trans*. In this model the cleavage is catalysed by cooperation of two enzymes: one bound to the single-site substrate and the other either free in solution (Figure 3.7A) or bound to another DNA (specific or non-specific) (Figure 3.7B). To examine the possible mode of single-site cleavage by EcoP1I, we carried out heparin trap assays and heterologous cooperation assays.

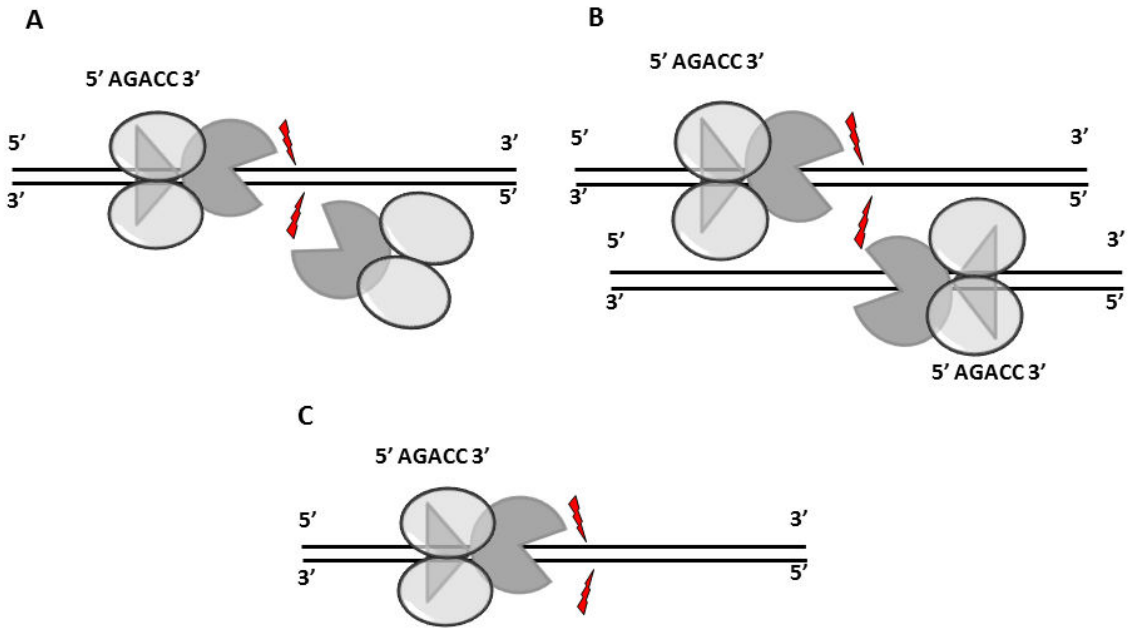


Figure 3.7: Possible interactions of two enzyme molecules to bring about dsDNA cleavage. The target (5' AGACC 3') site is shown as the triangle on the DNA. The grey closed ovals represent Mod1 and Mod2 while the oval with a mouth represents Res. A] *in trans:* cooperation between a DNA bound enzyme and a free enzyme molecule (from solution); B] *in trans:* cooperation between two DNA bound enzyme molecules; C] in *cis*: single enzyme molecule catalyzing a dsDNA break formation.

3.2.1 Heparin trap assays

To find if a free EcoP1I enzyme molecule in solution can cooperate with another molecule bound to single-site dsDNA in *trans*, we carried out a heparin trap assay utilizing 15/32_P1. Heparin is a linear polysaccharide, which is widely used to trap DNA binding proteins (22–25). The short length of the single-site DNA (52 bp) allowed binding of only one molecule of EcoP1I. When the enzyme dissociates from the DNA, excess heparin would trap it. This way, the possibility of cooperation/dimerization of a free and a DNA-bound enzyme was minimized, if not completely eliminated. To such a reaction mix, where certain percentage of enzyme molecules were bound to the single-site DNA and the rest were sequestered by heparin, ATP was added. Hydrolysis of ATP has been shown to promote EcoP1I to leave the target and move along the DNA (9). Due to the short length of the DNA, the enzyme would fall off immediately, and be sequestered by heparin, unless the conformational state of the enzyme is not proficient to bind heparin.

The heparin trap experiment was primarily done to sequester all free EcoP1I molecules in the solution. Upon sequestration of free EcoP1I molecules by heparin, any nucleolytic cleavage observed would be a result of one of the two possibilities:

1) a single enzyme bound to its target site eliciting a dsDNA break in *cis*;
2) or two DNA bound enzyme molecules cooperating to bring about dsDNA break.

We studied the sequestration of EcoP1I, pre-incubated with a fixed concentration of single-site dsDNA, in presence of varying concentrations of heparin. Increasing concentration of heparin reduced the binding of EcoP1I to DNA substrate suggesting a competition between the two polyanions. At heparin concentrations greater than 0.6 mg/mL, no EcoP1I-DNA complex or nucleolytic cleavage could be visualized (Figure 3.8A, Lanes 3-5). This suggested that at these concentrations heparin successfully competed with DNA to bind to most or all of EcoP1I. At reduced heparin concentrations, a small fraction of EcoP1I-DNA complex could be seen along with cleaved products (Figure 3.8A, Lanes 6-10). A titration with varying concentrations of heparin showed that 0.6 mg/mL heparin was sufficient to sequester all free EcoP1I, while still leaving some bound to DNA to catalyse DNA cleavage (Figure 3.8A, Lane 6 and Figure 3.8B). Increasing the incubation time of the reaction mix did not either increase or decrease the amount of cleaved DNA, indicating that the heparin-bound EcoP1I did not dissociate to bind and cleave the DNA.

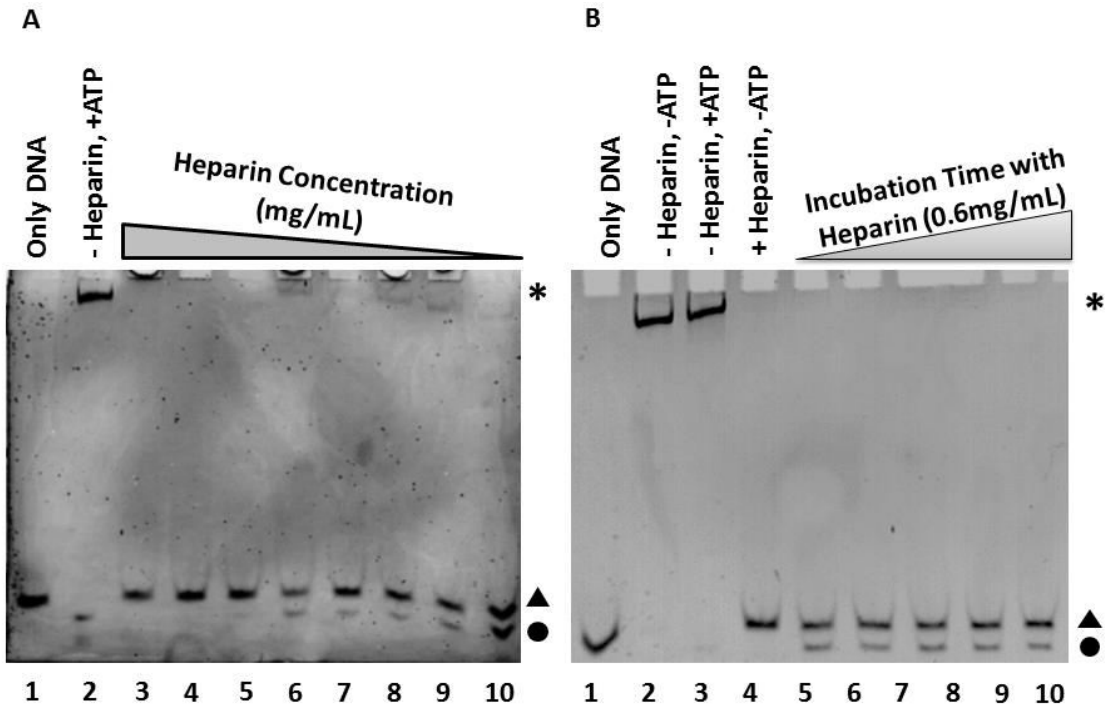


Figure 3.8: Effect of heparin. The reactions contained 500 nM EcoP1I and 200 nM oligomer in buffer D+. A] The effect of increasing concentration of heparin on binding of EcoP1I to target oligomer. 1)Only DNA, 2) –Hep, +ATP, Lanes 3-10) + Hep, +ATP: 3) Hep (3 mg/mL),4) Hep (2 mg/mL), 5) Hep (1 mg/mL), 6) Hep (0.6 mg/mL), 7) Hep (0.4 mg/mL), 8) Hep (0.3 mg/mL, 9) Hep (0.2 mg/mL), 10) Hep (0.1 mg/mL). B] Effect of increase in incubation time with heparin prior to ATP addition: 1) Only DNA 2) –Hep, -ATP 3) –Hep, +ATP 4) +Hep, -ATP , Lanes 5 to 10) depict 1, 2, 5, 10, 20, 40 minutes of incubation after addition heparin (0.6 mg/mL) and before addition of ATP respectively. * denotes EcoP1I-DNA complex, ▲denotes uncut DNA, ●denotes cleaved DNA.

Having established the sequestration of EcoP1I by heparin, we performed time dependent cleavage assays with and without heparin to assess the possible turnover of DNA bound EcoP1I molecules (Figure 3.9). When heparin was excluded from the reaction, almost 90% template DNA got cleaved in less than 15 minutes and the apparent initial rate of the reaction was 52 % DNA cleaved/min. When enzyme and heparin were incubated together with DNA no cleavage was seen even after 40 minutes (data not shown). This again demonstrated strong competition presented by heparin to DNA. When heparin was added to pre-incubated EcoP1I-DNA mix, however, the cleavage kick started at a much lower rate and the cleavage remained constant at 45% and the apparent initial rate of the reaction reduced to 5.6 % DNA cleaved/min. This amount of cleaved fraction remained constant over a long period of time, suggesting

that free EcoP1I molecules in solution were not responsible to bring about cooperation in *trans* with those bound to target. Hence, we concluded that the DNA-bound EcoP1I catalysed only a single round of DNA cleavage before being sequestered by heparin. We note that this assay does not eliminate the possibility of free enzymes acting in *trans* that are conformationally incompetent to bind to heparin.

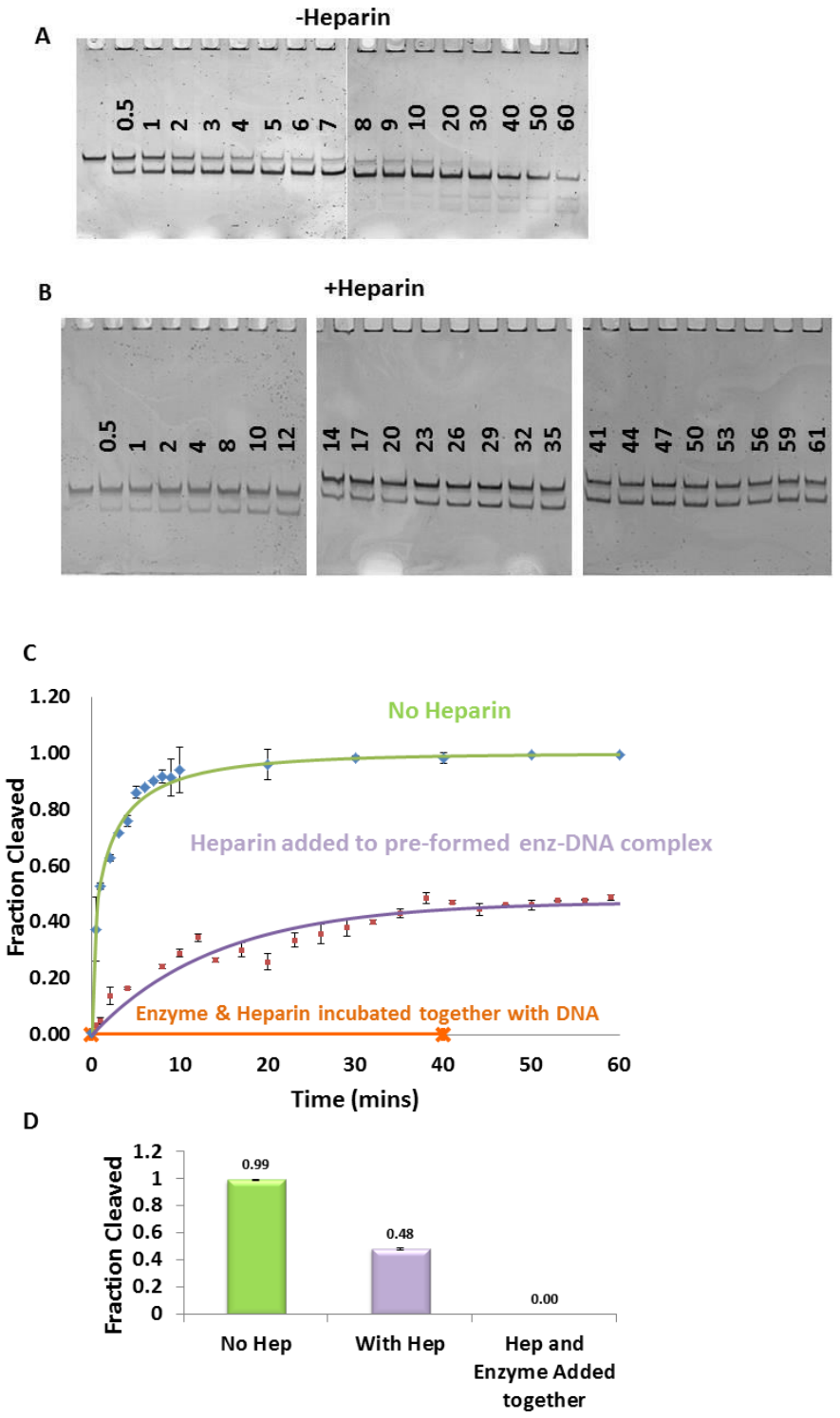


Figure 3.9: Effect of heparin (Hep) on cleavage kinetics. Reactions contained 500 nM EcoP1I and 300 nM 15/32_P1 in buffer D+. Time dependence of cleavage of 15/32_P1 by EcoP1I shown as electrophoretic separation of cleavage products in absence (A) and in presence (B) of heparin. Samples were withdrawn at time points (in minutes) indicated on gel. C] Cleavage kinetics of without heparin (green) and with heparin (purple). The orange line shows cleavage kinetics when enzyme and heparin were added together before addition of ATP; D] Bar graph representation of fraction of 15/32_P1 cleaved for each of these conditions

3.2.2 Heterologous cooperation assays reveal the cleavage to be a trans activity

We designed a 218 bp mixed HtH substrate containing one target site of EcoP1I and one of EcoP15I to check if the enzymes cooperated with each other to bring about dsDNA break (Figure 3.10). The enzymes cooperated with each other to cleave the substrate close to the target sites resulting in four fragments, ~180, 171, 47 and 38 bp (Figure 3.10, Lane 9). This is a *cis* event where both the co-operating enzymes are on the same piece of DNA. Interestingly, when one of the enzymes was omitted from the reaction, cleavage close to target site of the other was observed. When EcoP15I was omitted, cleavage close to EcoP1I target was observed and vice versa (Figure 3.10, Lanes 3-6).

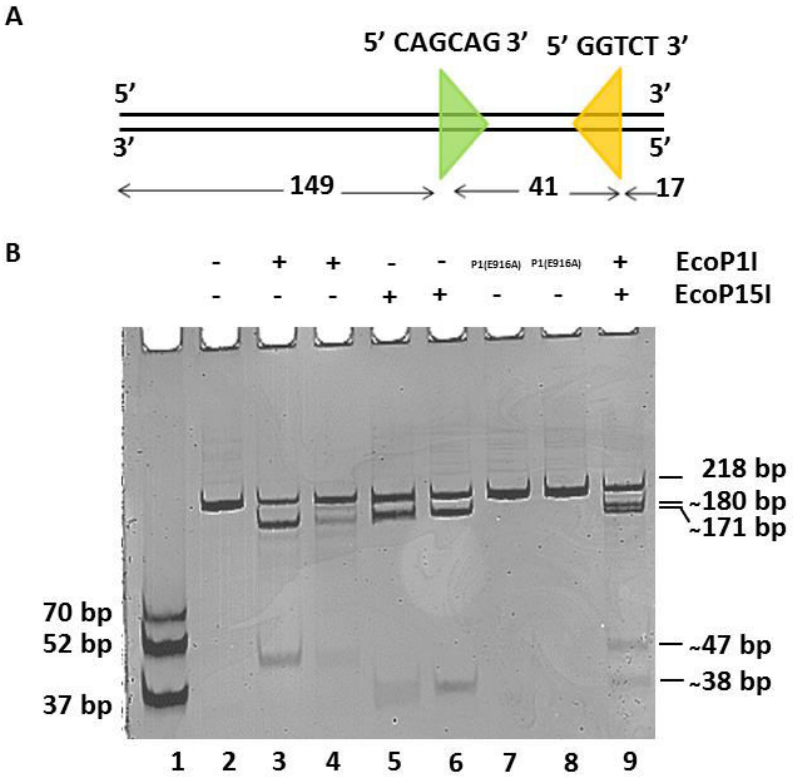


Figure 3.10: Cleavage assay with a mixed HtH substrate for EcoP1I and EcoP15I. A] Mixed HtH substrate with target sites for EcoP1I (yellow triangle) and EcoP15I (green triangle); B] Cleavage assay with mixed HtH substrate. The reactions contained 300 nM DNA in buffer D+. 1) Marker, 2) Control without enzyme, 3) 300 nM EcoP1I, 4) 500 nM EcoP1I, 5) 300 nM EcoP15I, 6) 500 nM EcoP15I, 7) 300 nM P1(E916A), 8) 600 nM P1(E916A), 9) 300 nM EcoP1I, 300 nM EcoP15I.

We used the above knowledge to design a "heterologous cooperation assay" to find if the single site cleavage is a *cis* or a *trans* activity. As the name suggests, the heterologous cooperation assay involved use of two separate short pieces of DNA each specific for EcoP1I or EcoP15I, respectively. The assay was carried out by adding ATP to a reaction mix containing EcoP15I and nuclease dead EcoP1I(E916A) bound to their respective single-site dsDNA. After 15 minutes, the reaction was stopped and the DNA in the reaction mix was analysed on a denaturing urea-formamide PAGE gel (Figure 3.11A). 5/24_EcoP15 was chosen as specific DNA for EcoP15I, since it did not undergo single-site dsDNA cleavage due to insufficient length downstream of the target site (data not shown).

The nuclease dead EcoP1I(E916A) did not show any endonucleolytic activity with 15/32_P1 (Figure 3.11A, Lane 2), while EcoP1I cleaved this oligomer resulting in two single strands of ~ 45 and 47 nucleotides (Figure 3.11A, Lane 3). Despite the use of the nuclease dead EcoP1I(E916A), the DNA containing its target was nicked, but not cleaved, in presence of EcoP15I bound to its specific DNA (Figure 3.11A, Lane 4). We also noted that EcoP15I did not nick EcoP1I specific DNA in absence of EcoP15I specific DNA (Figure 3.11A, Lanes 5, 7). Consequently, we concluded that the nicking of EcoP1I specific oligomer was a result of cooperation between EcoP1I bound to its specific DNA and EcoP15I bound to its specific DNA. The nick on the DNA specific to EcoP1I was catalysed by EcoP15I in *trans* only in presence of its specific DNA. However, our experiment did not rule out the possibility that the binding to the specific DNA is only required for activation of the enzyme to a reaction competent conformational state, and not actually during the reaction.

3.3 Insights into interaction of motor domain and DNA for successful trans interaction

We proceeded to further understand the relevance of the EcoP15I specific DNA in the above reactions. An earlier experiment showed that the EcoP15I specific DNA 5/24_P15 did not get cleaved suggesting that it is not long enough to reach the active site of the EcoP15I nuclease domain (data not shown). We next asked if there was a minimal length of specific DNA required for the *trans* activity to be observed. Towards this, we used EcoP15I specific DNA of varying lengths downstream of the target (Figure 3.11B). We saw nicking of the EcoP1I specific DNA even in the presence of 5/17_P15 (Figure 3.11C, Lane 6). However, with the downstream length of EcoP15I specific DNA shortening to 16 bp or less, the nicking of EcoP1I specific DNA was completely abolished (Figure 3.11C, Lanes 4, 5). This analysis suggested that for *trans* nicking activity of EcoP15I, the ATPase domain has to be engaged by downstream DNA and activated by ATP. From the partial structure of EcoP15I bound to a short DNA substrate mimic, having only 11 bp downstream of the target site (26), we see that the ATPase domain in not completely engaged with DNA. We propose that for successful interaction of ATPase domain and DNA, a minimum length of 17 bp is required downstream of the target site (Figure 3.11D).

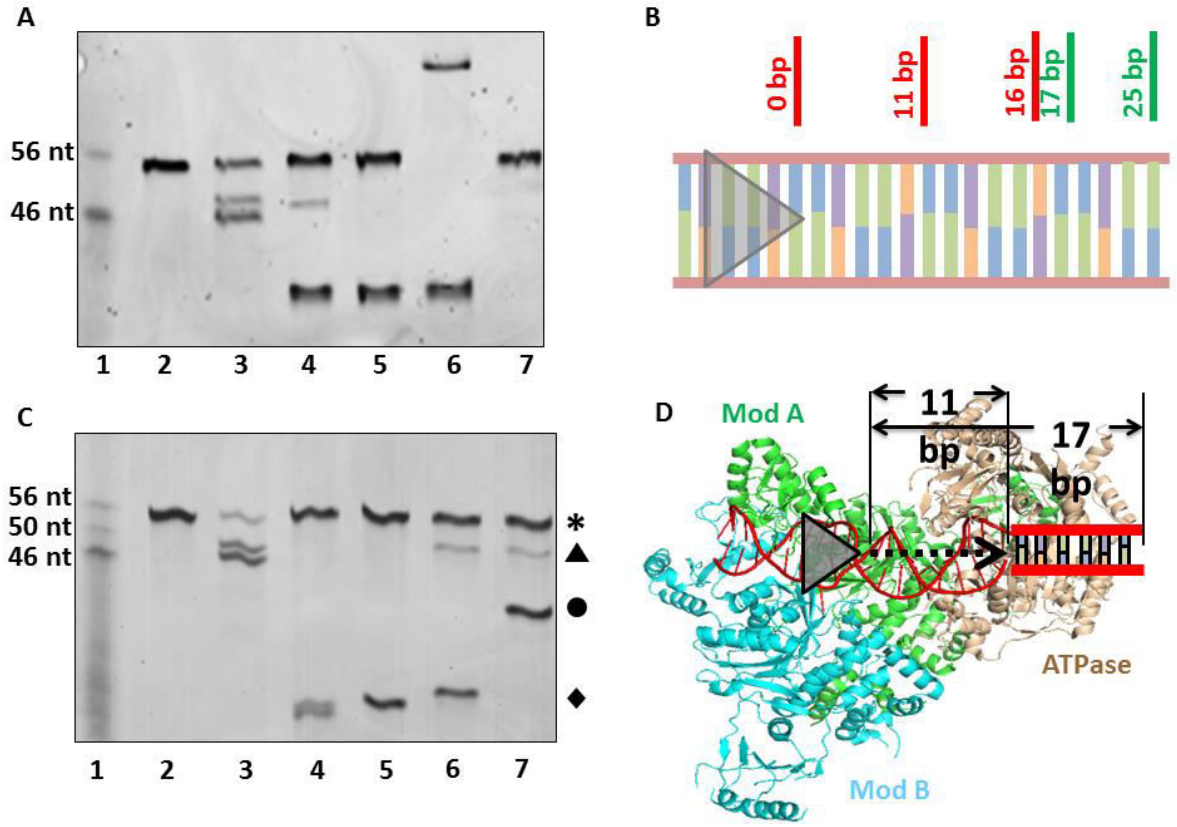


Figure 3.11: Heterologous cooperation assay: A] 1) Marker, 2) 15/32_P1 + EcoP1I (E916A), 3) 15/32_P1 + EcoP1I, 4) 15/32_P1+ EcoP1I (E916A) in cooperation with 5/24_P15 + EcoP15I, 5) 15/32_P1 + EcoP1I (E916A) in cooperation with NS_P15 + EcoP15I 6) NS_P1 + EcoP1I (E916A) in cooperation with 5/24_P15 + EcoP15I, 7) 15/32_P1 +EcoP15I. B] Schematic of variation of downstream length in specific oligomer for EcoP15I. The ladder represents dsDNA. The target site for EcoP15I (5' CAGCAG 3') is shown as triangle. The downstream length of DNA is numbered after last guanine (G) of the target site. C] Denaturing 15% Urea-Formamide Gel showing single stranded templates and products. 1) Marker, 2) 15/32_P1 + EcoP1IE916A, 3) 15/32_P1 + EcoP1I, 4) 15/32_P1+ EcoP1IE916A in cooperation with 3/11_P15 + EcoP15I, 5) 15/32_P1+ EcoP1IE916A in cooperation with 3/16_P15 + EcoP15I, 6) 15/32_P1+ EcoP1IE916A in cooperation with 3/17_P15 + EcoP15I, 7) 15/32_P1+ EcoP1IE916A in cooperation with 5/25_P15 + EcoP15I. * denotes specific DNA for EcoP1I, ▲ denotes nicked strands of 15/32_P1, ● and ♦ denote specific DNA for EcoP15I. Note that in lanes where EcoP1IE916A was used in cooperation with EcoP15I, only one of the strands of 15/32_P1 got nicked. D] Partial structure of EcoP15I with 3/11_P15 DNA shown in red. Mod_A: Green, Mod_B: Cyan, Res: Brown. The target site for EcoP15I (5' CAGCAG 3') is shown as a grey triangle. The extended DNA upto 17 bp downstream of target is shown as a ladder.

4. Discussion

To cleave the dsDNA, Type III RM enzymes need to communicate between two distantly located target sites (1, 3, 9, 14, 27–29). Moreover, the target sites need to be inversely oriented to achieve successful interaction of two enzymes, whose orientation is dictated by that of target site (3, 27). To explain long range communication and site orientation selectivity associated with Type III RM enzymes, a number of models have been proposed (3, 4, 6, 7, 9, 28, 30). However none of these modes of communication could explain nucleolytic cleavage of DNA substrates bearing single target site. Nucleolytic cleavage of single-site DNA substrates was observed earlier by a few researchers (14, 16, 18, 19, 31). However, it was overlooked grossly. A couple of studies also reported cleavage of single-site oligomers as short as 70 bp and 50 bp by EcoP1I and EcoP15I (14, 19). Current modes of communication proposed for Type III RM enzymes fail to explain such endonucleolytic activity, since juxtaposition of two enzymes cannot be achieved on such DNA substrates due to spatial constraints. This is discussed in section 3.2.

Characterization of nucleolytic cleavage of single-site DNA : We investigated an unusual mode of cleavage by Type III RM enzyme, EcoP1I, where it cleaved a substrate with single target site. After characterizing this phenomenon, we found out that the properties of cleavage were similar to that of canonical cleavage by Type III RM enzymes. The cleavage took place 26 nucleotides downstream (confirmed by sequencing the cleaved product of 1324/1364_P1) and 24 nucleotides downstream for a family of short oligomers (confirmed by monitoring the cleavage products on a denaturing PAGE gel). Similar to cleavage product of substrates with inversely oriented sites, we saw an overhang of approximately 2 nucleotides in the cleavage product of single-site DNA substrate. The cleavage took place only with a specific DNA with a strict requirement of ATP. Mere nucleotide binding was insufficient to trigger the nucleolytic activity since cleavage did not take place in presence ADP, AMP-PNP or transition state ATP analogue ADP-vanadate.

In addition, cleavage was observed irrespective of length DNA substrate. As mentioned earlier, single site cleavage by EcoP1I and EcoP15I was reported in literature, however, groups working with longer DNA substrates observed that the cleavage was weak (16, 18, 31). In reports, where shorter oligomers were used for assays, the cleavage seemed

efficient (14, 19). Consistent with these reports, we observed that the efficiency of cleavage depended on the length of DNA. The extent of cleavage was highest with shorter oligomers and decreased with longer DNA substrates. This effect cannot be attributed to length of the DNA alone, as the reaction mix with short DNA had more amounts of both enzyme and DNA (The detection method was staining by ethidium bromide, and hence more amount of shorter DNA was required to visualize on gel). As the crowding of enzyme-DNA complex in the reaction mix increased from long to short DNA, it could have an effect on overall cleavage efficiency. Later we demonstrated that this type of cleavage occurs by *trans* interaction of two enzyme-DNA complexes. The crowding of enzyme-DNA complexes leading to increase in cleavage efficiency of single-site substrates is a direct consequence of such a *trans* interaction. In conclusion, EcoP1I cleaved single-site substrates of all the lengths with cleavage efficiencies appearing higher for shorter oligomers.

Further we sought to understand the effect of AdoMet and SF on cleavage efficiency in presence of potassium or sodium ions. There are contradictory reports on requirement of AdoMet for cleavage by Type III RM enzymes (16, 17, 32). The role of AdoMet and sodium ions in preventing the promiscuous cleavage by EcoP1I was reported earlier by Peakman et al (16). It was also suggested that potassium ions leads to promiscuous enzymatic activity (16). Here, the authors termed any cleavage event taking place on a substrate containing directly repeated sites and/or single site as secondary or promiscuous cleavage. Also, Raghavendra et al have reported the effect of SF on the cleavage by EcoP15I (17). SF is a structural analogue and competitor of AdoMet that lacks a transferrable methyl group to target adenine. We observed that nucleolytic cleavage of single-site substrate DNA occurred irrespective of presence of AdoMet or SF in the reaction mix. In presence of potassium ions, cleavage reaction was efficient; however, a lot of extra bands resulting from non-specific cleavage events were observed. The presence of SF increased specificity of the reaction. In presence of sodium ions, an overall decrease in cleavage efficiency was observed which improved upon addition of SF. AdoMet had an inhibitory effect on the cleavage possibly due to methylation of target site. We know that the MTase of Type III RM systems recognizes and methylates the target adenine in the target sequence (12, 13). The methylated site is no longer identified by the enzyme as a target for activating endonuclease activity.

Protection from cleavage upon methylation provided an indirect evidence for the single-site cleavage to be site specific.

The unmodified single-site substrates stimulated the ATPase activity of EcoP1I. This is consistent with the ATP-dependent nucleolytic activity of the enzyme and also a previous study characterising the ATPase activity of Type III RM. This ATPase activity reduced by 10 fold when the single-site substrate was methylated specific, which is comparable to the ATPase activity of the enzyme in presence of non-specific oligomers. Surprisingly, we found that when ATP and AdoMet were added together to EcoP1I pre-incubated with specific oligomer, the ATPase activity was poor and comparable to that noticed with methylated substrate. This suggested that the rate of methylation is faster than that of ATP hydrolysis by Type III RM enzymes. To the best of my knowledge, this is the first observation on comparison of rates of methylation and ATP hydrolysis associated with single-site substrates. Earlier, Muecke et al demonstrated that DNAse footprinting assays with single-site DNA in simultaneous presence of ATP and AdoMet resulted in very weak footprint of EcoP15I on substrate DNA as compared to that with ATP or AdoMet alone (14). The authors suggested that methylation reduced the DNA affinity of EcoP15I in presence of ATP. These rates, however, are calculated utilizing a non-radioactive procedure (using malachite green) to detect inorganic phosphate. Further investigations utilizing labelled elements in AdoMet and ATP to quantify these rates would result in accurate estimation and interpretation of such a comparison.

To summarize, the properties of nucleolytic cleavage of single-site DNA substrates by EcoP1I were demonstrated to be similar to that of DNA substrates with inversely oriented sites.

Cleavage of single-site DNA substrates by EcoP1I/15I is a trans event: Type III RM enzymes identify a pair of indirectly repeated unmethylated target sites through Mod, which in turn activates Res to trigger ATP hydrolysis resulting in dsDNA break 25-27 bp downstream of the target site. It is demonstrated that such a dsDNA break results from a pair of nicks, where the top and bottom strand nick are contributed by proximal and distal enzymes respectively (15). Also, using DNA substrates where the target sites for EcoP1I were on two different rings of catenane, Peakman et al showed that there is a

strict requirement of two target sites in *cis* for successful interaction between two macromolecular assemblies (8). In order to understand mechanism behind nucleolytic cleavage of single-site DNA substrates, we performed experiments with long DNA substrates (>1000 kb) as well as extremely short oligomers (<40 bp). Upon systematically monitoring the cleavage of various oligomers, we noticed that the cleavage ceased only when the oligomer had < 25 bp downstream of the target site. Thus a downstream length of just 25 bp was sufficient and necessary for a successful dsDNA break. It is hard to envisage simultaneous occurrence of two heterotrimeric EcoP1I/P15I molecules on such a short DNA piece. Observations from DNAase footprinting assays (14), SEC-MALS analysis (19) and partial crystal stricture of EcoP15I bound to substrate DNA mimic (26) further suggest that only single heterotrimeric assembly of Mod_2Res can bind such a short oligomer. We considered three possibilities of occurrence of nucleolytic cleavage of single-site DNA substrates as discussed in section 3.2: 1) *in trans:* cooperation between a DNA bound enzyme and a free enzyme molecule (from solution), 2) *in trans:* cooperation between two DNA bound enzyme molecules and 3) *in cis*: single enzyme molecule executing a dsDNA break. To find out whether free enzyme molecules in solution could cooperate with target site bound enzyme molecule, we performed heparin trap experiments. The results of these experiments suggested that nucleolytic cleavage of single-site DNA is a single turnover reaction and that there is very less possibility of free enzymes (in solution) to cooperate with target site bound enzymes. Further we designed "heterologous cooperation assay" to directly probe the interaction between two enzyme-DNA complexes. It has been shown earlier that EcoP1I and EcoP15I cooperate with each other in *cis* to cleave DNA containing one target of each of the enzymes in Head-to-Head (HtH) orientation (33). We conceptualized a scenario where a target bound EcoP1I and EcoP15I could juxtapose such that the enzymes are oriented in a HtH fashion. When one of these enzymes was replaced by a mutant lacking endonuclease activity (EcoP1I(E916A) in our case), we could monitor the nick on EcoP1I substrate contributed by other enzyme (EcoP15I). This nick was a result of *trans* activity. To our surprise, EcoP15I nicked the EcoP1I specific oligomer in *trans.* In addition, this nicking showed a strict requirement of specific substrates for enzymes (EcoP1I and EcoP15I). EcoP15I in presence of a non-specific substrate or devoid of its specific substrate could not cooperate with target bound EcoP1I resulting in abolishment of nicking. Similarly, EcoP15I could not

cooperate in conditions where EcoP1I specific oligomer was excluded from the reaction. These results show that nucleolytic cleavage of single-site DNA substrates is a result of *trans* cooperation between two target bound enzyme molecules. Such a cleavage mode has also been suggested for a Type IIGS restriction enzyme Tth111II (34).

Insights on requirement of sufficient downstream length of DNA to engage the ATPase domain: From the partial structure of EcoP15I bound to a short DNA substrate mimic (26), we see that the ATPase domain in not completely engaged with DNA. Gupta et al crystallized EcoP15I with a DNA substrate mimic which has only 11 bp downstream of the target site. We used heterologous cooperation assay to examine necessary length of DNA downstream of target site for latching of ATPase domain onto polynucleotide track. By systematically altering the length of DNA downstream of target site, we observed that EcoP15I failed to cooperate with EcoP1I in presence of specific oligomers with <17bp downstream of target site. With our studies, we propose that minimum 17 bp is required for successful interaction between DNA and the ATPase domain. This information was vital to design an appropriate DNA substrate mimic suitable for crystallization of EcoP1I-DNA complex.

In summary, we demonstrated that Type III RM enzymes EcoP1I and EcoP15I can communicate with two target sites by a 3D route, where two target site associated and ATP-activated enzyme molecules cooperate with each other in *trans* to carry out dsDNA break by contribution of single nick per enzyme molecule. Additional experiments are required to be done to obtain a detailed molecular mechanism of this *trans* nucleolytic activity. We need to investigate on the requirement of motor activity for either *cis* or *trans* acting or both enzymes for successful nicking of each DNA strand. Heterologous cooperation assays with EcoP1I and EcoP15I where key ATPase motifs are mutated could answer this question. In addition, we still do not know whether the enzyme molecule interacting in *trans* needs to be bound to site; or an ATP activated sliding enzyme which is still bound to DNA non-specifically is also able to cooperate.

In conclusion, we have learnt a new three dimensional mode of communication between 2 target sites by Type III RM enzymes.

5. References

1. Saha,S. and Rao,D.N. (1995) ATP hydrolysis is required for DNA cleavage by EcoPI restriction enzyme. *J. Mol. Biol.*, **247**, 559–567.

2. Toth,J.,Bollins,J. and Szczelkun,M.D. (2015) Re-evaluating the kinetics of ATP hydrolysis during initiation of DNA sliding by Type III restriction enzymes. *Nucleic Acids Res.*, **43**, 10870–10881.

3. Aelst,K. Van, Tóth,J., Ramanathan,S.P., Schwarz,F.W., Seidel,R. and Szczelkun,M.D. (2010) Type III restriction enzymes cleave DNA by long-range interaction between sites in both head-to-head and tail-to-tail inverted repeat. *PNAS*, **107**, 9123–9128.

4. Ramanathan,S.P., Aelst,K. Van, Sears,A., Peakman,L.J., Diffin,F.M., Szczelkun,M.D. and Seidel,R. (2009) Type III restriction enzymes communicate in 1D without looping between their target sites. *PNAS*, **106**, 1748–1753.

5. Dryden,D.T.F., Murray,N.E. and Rao,D.N. (2001) Nucleoside triphosphate-dependent restriction enzymes. *Nucleic Acids Res.*, **29**, 3728–3741.

6. Raghavendra,N.K. and Rao,D.N. (2004) Unidirectional translocation from recognition site and a necessary interaction with DNA end for cleavage by Type III restriction enzyme. *Nucleic Acids Res.*, **32**, 5703–5711.

7. Crampton,N., Yokokawa,M., Dryden,D.T.F., Edwardson,J.M., Rao,D.N., Takeyasu,K., Yoshimura,S.H. and Henderson,R.M. (2007) Fast-scan atomic force microscopy reveals that the type III restriction enzyme EcoP15I is capable of DNA translocation and looping. *PNAS*, **104**, 12755–12760.

8. Peakman,L.J. and Szczelkun,M.D. (2004) DNA communications by Type III restriction endonucleases - confirmation of 1D translocation over 3D looping. *Nucleic Acids Res.*,**32**, 4166–4174.

9. Schwarz,F.W., Toth,J., van Aelst,K., Cui,G., Clausing,S., Szczelkun,M.D. and Seidel,R. (2013) The helicase-like domains of type III restriction enzymes trigger long-range diffusion along DNA. *Science.*, **340**, 353–356.

10. Gorman,J., Wang,F., Redding,S., Plys,A.J., Fazio,T., Wind,S., Alani,E.E. and Greene,E.C. (2012) Single-molecule imaging reveals target-search mechanisms during DNA

mismatch repair. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, E3074-83.

11. Kelch,B.A., Makino,D.L., O'Donnell,M. and Kuriyan,J. (2011) How a DNA polymerase clamp loader opens a sliding clamp. *Science.*, **334**, 1675–1680.

12. Hadi,S.M., Baechi,B., Shepherd,J.C.W., Yuan,R., Ineichen,K. and Bickle,T.A. (1979) DNA recognition and cleavage by the EcoPl5 restriction endonuclease. *J. Mol. Biol.*, **134**, 655–666.

13. Baechi,B., Reiser,J. and Pirrota,V. (1979) Methylation and cleavage sequences of the EcoPl enzyme. *J. Mol. Biol.*, **128**, 143–163.

14. Merlind,M., Reich,S., Mo,E., Reuter,M. and Kruger,D.H. (2001) DNA cleavage by Type III restriction-modification enzyme EcoP15I is independent of spacer distance between two head to head oriented recognition. *J. Mol. Biol.*, **312**, 687–698.

15. Janscak,P., Sandmeier,U., Szczelkun,M.D. and Bickle,T.A. (2001) Subunit assembly and mode of DNA cleavage of the type III restriction endonucleases EcoP1I and EcoP15I. *J. Mol. Biol.*, **306**, 417–431.

16. Peakman,L.J., Antognozzi,M., Bickle,T.A., Janscak,P. and Szczelkun,M.D. (2003) S-Adenosyl methionine prevents promiscuous DNA cleavage by the EcoP1I type III restriction enzyme. *J. Mol. Biol.*, **333**, 321–335.

17. Raghavendra,N.K. and Rao,D.N. (2005) Exogenous AdoMet and its analogue sinefungin differentially influence DNA cleavage by R.EcoP15I—Usefulness in SAGE. *Biochem. Biophys. Res. Commun.*, **334**, 803–811.

18. Peakman,L.J. and Szczelkun,M.D. (2009) S-Adenosyl homocysteine and DNA ends stimulate promiscuous nuclease activities in the Type III restriction endonuclease EcoPI. *Nucleic Acids Res.*, **37**, 3934–3945.

19. Butterer,A., Pernstich,C., Smith,R.M., Sobott,F., Szczelkun,M.D. and Tóth,J. (2014) Type III restriction endonucleases are heterotrimeric: comprising one helicase-nuclease subunit and a dimeric methyltransferase that binds only one specific DNA. *Nucleic Acids Res.*, **42**, 5139–50.

20. Geladopoulos,T.P., Sotiroudis,T.G. and Evangelopoulos,A.E. (1991) A malachite green colorimetric assay for protein phosphatase activity. *Anal. Biochem.*, **192**, 112–116.

21. Baykov,A.A., Evtushenko,O.A. and Avaeva,S.M. (1988) A malachite green procedure for orthophosphate determination and its use in alkaline phosphatase-based enzyme immunoassay. *Anal. Biochem.*, **171**, 266–270.

22. King,A.J., Teertstra,W.R., Blanco,L., Salas,M. and Van Der Vliet,P.C. (1997) Processive proofreading by the adenovirus DNA polymerase. Association with the priming protein reduces exonucleolytic degradation. *Nucleic Acids Res.*, **25**, 1745–1752.

23. Subramanian,K., Rutvisuttinunt,W., Scott,W. and Myers,R.S. (2003) The enzymatic basis of processivity in λ exonuclease. *Nucleic Acids Res.*, **31**, 1585–1596.

24. Fischer,C.J., Maluf,N.K. and Lohman,T.M. (2004) Mechanism of ATP-dependent translocation of E.coli UvrD monomers along single-stranded DNA. *J. Mol. Biol.*, **344**, 1287–1309.

25. Hogg,M., Osterman,P., Bylund,G.O., Ganai,R. a, Lundström,E.-B., Sauer-Eriksson, a E. and Johansson,E. (2014) Structural basis for processive DNA synthesis by yeast DNA polymerase ε. *Nat. Struct. Mol. Biol.*, **21**, 49–55.

26. Gupta,Y.K., Chan,S.-H., Xu,S. and Aggarwal,A.K. (2015) Structural basis of asymmetric DNA methylation and ATP-triggered long-range diffusion by EcoP15I. *Nat. Commun.*, **6**, 1–10.

27. Meisel,A., Bicklel,T.A., Kruger,D.H. and Schroeder,C. (1992) Type III restriction enzymes need two inversely oriented recognition sites for DNA cleavage. *Nature*, **355**, 467–469.

28. Kruger,D.H., Kupper,D., Meisel,A., Tierlich,M., Reuter,M. and Schroeder,C. (1995) Restriction endonucleases functionally interacting with two DNA sites. *Gene*, **157**, 165.

29. Sears,A., Peakman,L.J., Wilson,G.G. and Szczelkun,M.D. (2005) Characterization of the Type III restriction endonuclease PstII from Providencia stuartii. *Nucleic Acids Res.*, **33**, 4775–4787.

30. Crampton,N., Roes,S., Dryden,D.T.F., Rao,D.N., Edwardson,J.M. and Henderson,R.M. (2007) DNA looping and translocation provide an optimal cleavage mechanism for the type III restriction enzymes. *EMBO J.*, **26**, 3815–3825.

31. Möncke-buchner,E., Rothenberg,M., Reich,S., Wagenführ,K., Matsumura,H., Terauchi,R., Krüger,D.H. and Reuter,M. (2009) Functional characterization and modulation of the DNA cleavage efficiency of Type III restriction endonuclease EcoP15I in its interaction with two sites in the DNA target. *J. Mol. Biol.*, **387**, 1309–1319.

32. Sistla,S. and Rao,D.N. (2004) S-adenosyl-L-methionine – dependent restriction enzymes. *Crit. Rev. Biochem. Mol. Biol.*, **39**, 1–19.

33. Kunz,A., Mackeldanz,P., Mucke,M., Meisel,A., Reuter,M., Schroeder,C. and Kruger,D.H. (1998) Mutual activation of two restriction endonucleases: interaction of EcoP1 and EcoP15. *Biol. Chem.*, **379**, 617–620.

34. Zhu,Z., Guan,S., Robinson,D., Fezzazi,H. El, Quimby,A. and Xu,S. (2014) Characterization of cleavage intermediate and star sites of. *Nat. Sci. Reports*, 10.1038/srep03838.

CHAPTER 4

Sequence and Structure Analyses of Type III RM Enzymes

Chapter 4

Sequence and Structure Analyses of Type III RM Enzymes

1. Introduction

Primary amino acid sequence forms the basic framework of the protein. Information on conserved motifs, functional modules and sequence based phyletic relation between different proteins can be obtained when the primary sequence of a large number of homologous proteins performing similar function are analyzed together. A frequently used and powerful means for such analyses is Multiple Sequence Alignment (MSA) of the amino acid sequences, where sequences are aligned optimally by bringing the maximum number of similar residues for a given position in the same column of the alignment. A carefully built MSA can yield information about conserved functional domains, conserved key motifs, insertions, deletions and evolutionary relationships between different members of the alignment. Furthermore, if structure of any of the members of alignment is known, we can predict spatial relationships of amino acids in other proteins in the alignment with the help of structure based sequence alignment (1–3). Conserved stretches of amino acids, generally referred to as motifs, is often suggestive of their functional or structural importance. A motif could be vital for interaction with a specific ligand or a co-factor, dimerization between subunits or could just have implications in maintaining the tertiary structure of the protein. MSA also reveals the presence of insertions in a particular family of proteins, which serve as additional functional modules (4). Thus, MSA forms an important tool to identify putative features of a family of proteins and enables us to predict possible functional role.

We initiated our research project on the Type III RM enzymes by carrying out a detailed study of their amino acid sequences. As structural information of these enzymes were completely absent when we initiated the study, the aim was to a) locate/identify the various functional motifs conserved in this type of enzymes; b) locate and delineate the boundaries of various functional and structural domains. As described in earlier chapters, Type III RM enzymes are composed of three subunits, two copies of the methylating Mod subunit (Mod_A and Mod_B) and a single copy of the endonucleolytic Res subunit (5–7). Each subunit, in turn, is made up of different functional modules. Mod has a Target Recognition Domain (TRD) inserted between the AdoMet binding and

catalytic domains of the MTase. The Mod_A and Mod_B subunits exhibit asymmetry in function, where Mod_A is responsible for identification of specific target site and Mod_B is involved in methylation of the target adenine (5). Res is a fusion of an ATPase and endonuclease domains (3, 8, 9). The fusion of various functional domains is also seen in other ATP dependent RM enzymes, viz. Type I and Type ISP, both of which restrict unmodified DNA by carrying out ATP dependent dsDNA translocation.

In our endeavor to gain a better picture of domain architecture, comparison of the amino acid sequences of a large number of Type III RM enzymes was carried out. In 2015, a partial structure of EcoP15I co-crystallized with AMP and a DNA substrate mimic determined by the laboratory of Aneel Aggarwal at Mount Sinai Hospital was available (5). The information gained by us from the sequence analysis of Type III RM enzymes were mapped onto the three dimensional structure of EcoP15I in order to get insights into their structural and functional relevance.

These analyses have led to identification of certain unique features of Type III RM enzymes which were previously unknown. We identified six new motifs in the MTase domain of the Type III enzymes. Out of these six motifs, five are involved in dimerization of the two Mods and one is involved in AdoMet binding. The analyses also revealed that the sequence length of Mod was variable. This variation affected the length of the C-terminal domain (CTD) while the length of other functional domains was less variable, leading to certain functional implications discussed below. We also identified a new motif unique to the ATPase domain of Type III RM enzymes, which is involved in recognizing ATP. The sequence analysis of the endonuclease domain of Type III RM enzymes led to identification of motif I.

In conclusion, sequence analyses empowered us to understand the enzyme system better, and also led to identification of novel sequence features.

2. Materials and Methods

2.1 Multiple sequence alignments

2.1.1 Domain-wise alignments for phylogenetic analyses (MTase, ATPase, endonuclease)

Homologous sequences were searched using EcoP1I as the query sequence using the database search engine BLASTP (10) against REBASE (11) by setting the sequence identity window to 20-95%. This analysis was carried out using sequences in REBASE as on 25-11-2010. Only a representative sequence was used for a set of sequences having more than 95% sequence identity amongst them. This helped to remove the redundant sequences (sequence identity greater than or equal to 95%). This was done in order to obtain a diverse set of sequences. The full length sequences of Res were delineated into ATPase and endonuclease domains based on previous information about domain boundaries (8, 12, 13). Multiple sequence alignments were done separately for each domain in ClustalX2 stand-alone tool (14) with default parameters. Each alignment step was iterated during complete alignment procedure. The resulting alignments were manually edited and polished using SeaView (15) and JalView (16). Sequence logos were generated using JalView and WebLogo online application (16, 17). Evolutionary conservation scores were calculated using ConSurf (18) for Type III MTase. Structure based MSA were generated using PROMALS 3D (multiple sequence and structure alignment server) (19).

2.1.2 MSA of Archeal Holliday Junction Resolvases (AHJR)

Homologous sequences were searched using a Holliday junction resolvase from *Pyrococcus furiosus* (PDB ID: 1GEF) as the query sequence for BLASTP (10) against non-redundant (nr) protein database (NCBI). First 100 sequences were selected for further analysis with sequence identity window of 35-96%. Redundant sequences were not removed since this set of sequences was used to see the known canonical motifs within AHJR family of nucleases. Subsequently MSA was done exactly as described in Section 2.1.1. Sequence logos were generated using JalView and WebLogo online application (16, 17).

2.2 Phylogenetic analyses

The multiple sequence alignment generated for each domain of Type III RM enzymes was used as input for inferring phylogeny. The phylogenetic trees were generated using Phylip stand-alone tool (20). The amino acid sequences of EcoR124I (a representative of Type I RM enzymes) and LlaGI (a representative of Type ISP RM enzymes) were used as outgroup sequences to obtain a rooted tree. First, a distance matrix was generated from the multiple sequence alignment using the program PROTDIST. The distance matrix was used as input for generating a NJ (Neighbor Joining) tree using the program NEIGHBOR. The resulting data set was resampled using SEQBOOT and a final tree was generated using CONSENSE. The phylogenetic tree was visualized using Dendroscope 3 (21).

2.3 Mapping the sequence features on structure

Pairwise structural alignments were done using DaliLite-pairwise option version 3.1 (22). PyMOL was used to visualize protein structures (23). The outputs from ConSurf (18) were mapped on to the structure using Chimera (24) for Type III methyltransferases.

2.4 MTases for length comparison

A complete dataset of 1061 sequences of Type III MTases was downloaded from REBASE on 11-9-2012. No sequence identity cut-off was set prior to MSA. MSA was done exactly as described in Section 2.1.1. The lengths of N-terminal domain (NTD), catalytic domain, TRD, AdoMet binding domain and CTD for each sequence were extracted from the alignment using JalView (16) by delineating domains using information from aligned canonical motifs as follows: From N-terminus to motif IV: NTD, motif IV- VIII: catalytic domain, motif VIII-motif X: TRD, motif X-motif III: AdoMet binding domain, beyond motif III: CTD.

3. Results

The sequence analysis was done separately for each functional module i.e. MTase, ATPase and endonuclease. The following section describes results obtained from these analyses.

3.1 Unique features of MTase of Type III RM enzymes

The MTases of Type III RM enzymes have TRD inserted between the N-terminal AdoMet binding and C-terminal catalytic domains (Figure 1.4A, Chapter 1). This linear arrangement of domains puts these enzymes under ß class of MTases according to the classification proposed by Malone et al (3). To identify unique features of Type III MTases, we compared 92 sequences of MTases of Type III RM enzymes. After exhaustive sequence alignments, a rooted NJ phylogenetic tree was constructed, to examine if sequences cluster due to similarities or differences at primary sequence level.

3.1.1 Identification of non-canonical motifs

The phylogenetic tree revealed that the 92 sequences clustered into 3 distinct groups while a few sequences were outliers such as CthVORF480P and LheDORF27P (Figure 4.1). Most of the canonical motifs of the ß MTase were conserved in all the three groups. A careful inspection of the MSA revealed certain motifs unique to each of the three groups. These conserved stretches of amino acids were different from the canonical motifs of ß MTase core, and hence we referred to them as non-canonical motifs. These motifs were found at the N-terminus of the MTase ahead of the canonical motif IV (Figure 4.2). One of these motifs, glycine-aspartate-asparagine (GDN), was found conserved in all the three groups. In Group I, two other non-canonical motifs were found in the catalytic domain.

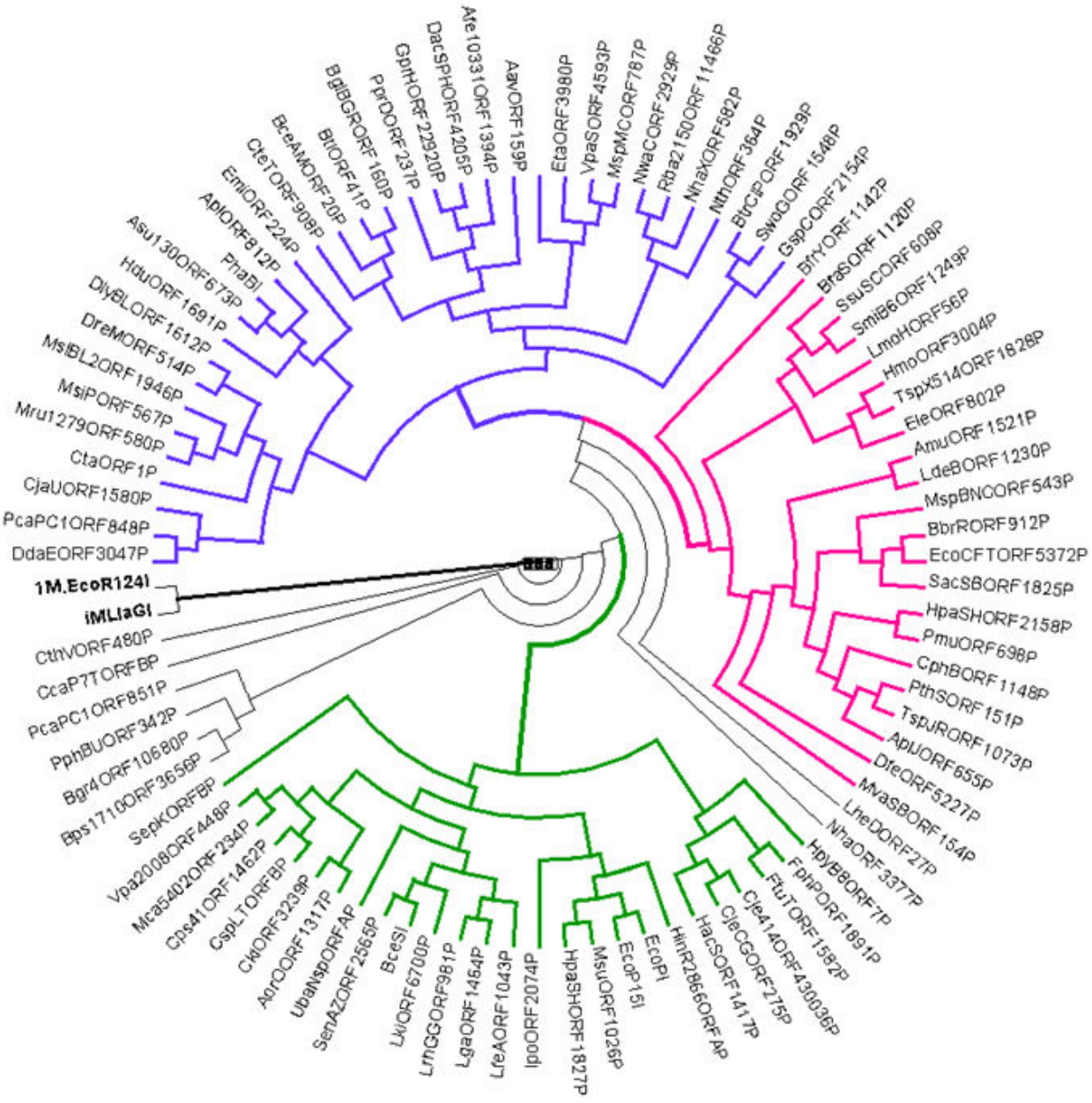


Figure 4.1: Phylogenetic analyses of MTases of Type III RM enzymes. A] Rooted NJ phylogenetic tree of MTases of Type III RM enzymes. MTases of EcoR124I (Type I RM enzyme) and LlaGI (Type ISP RM enzyme) were used as an outgroup and is shown by a thick black line. The three groups are highlighted as green (Group I), pink (Group II) and blue (Group III). The outliers are shown in black lines.

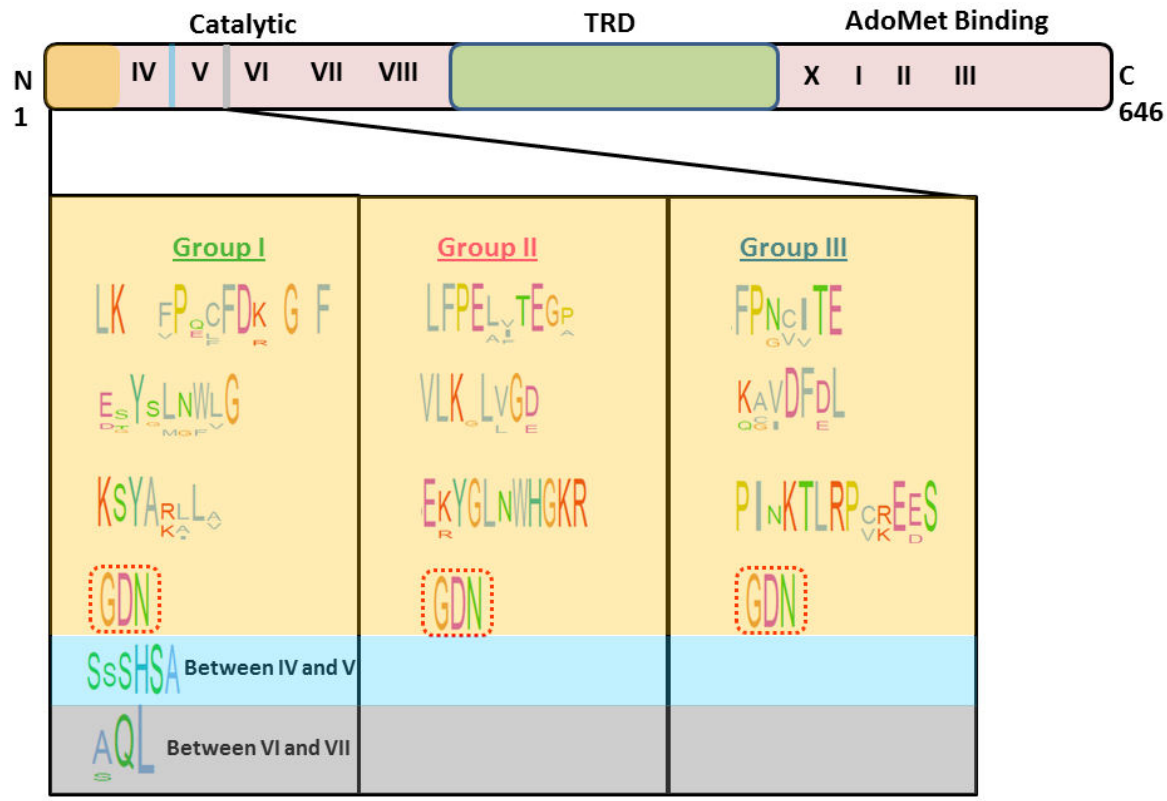


Figure 4.2: Unique conserved motifs of MTases of Type III RM enzymes. The Mod subunit is represented as a combination of AdoMet binding domain, catalytic domains (pink blocks) and (TRD) (green block). Roman numerals represent canonical motifs within ß MTases. The location of non-canonical motifs is highlighted by orange, blue and grey rectangles. The height of letters corresponds to the degree of conservation. The conserved glycine-aspartate-asparagine (GDN) is found in all the three groups, hence is highlighted by red serrated rectangle.

The sequence-based comparison of Type III MTases led to identification of previously undefined non-canonical motifs. To gain insights into their functional relevance we mapped the residues on to the structure of EcoP15I, which belongs to Group I. We observed that all the motifs except GDN are a part of the dimeric interface between Mod_A and Mod_B (Table 4.1). Along with the motifs at the N terminus, two other motifs were identified in the catalytic domain of the MTase (Figure 4.2). Structural mapping of the two motifs revealed them to be involved in Mod_A-Mod_B dimerization rather than catalysis (Table 4.1).

As mentioned earlier, the motif GDN was found conserved in all the three groups (Figure 4.2). This motif was found to be located in the AdoMet binding pocket. To examine if this motif interacted with AdoMet, we compared the structures of two stand-

alone ß MTases (MTases that are functionally competent on their own), M1.MboII (PDB ID: 1G60) and M.RsrI (PDB ID: 1NW5) co-crystallized with AdoMet with that of EcoP15I (PDB ID: 4ZCF). The conserved aspartate D101 of GDN plays a dual role by interacting via a hydrogen bond with the exocyclic nitrogen N6 of AdoMet (Figure 4.3B) and capping the helix spanning N102 to Y113. We observed that an aspartate (D46) in M.RsrI and an asparagine (N11) in M1.MboII at equivalent positions interact with N6 of AdoMet and cap the helices C47-L53 (M.RsrI) and C12-V19 (M1.MboII), respectively (Figure 4.3A). The superposition of these residues can be seen in structure based sequence alignment of M1.MboII, M.RsrI and EcoP15I (Figure 4.4A). The sequence analysis of Type III RM enzymes showed that the aspartate interacting with AdoMet is strongly conserved in Mod. Hence, the triplet GDN is a motif unique to Type III RM enzymes.

In summary, we identified 6 new motifs associated with MTases of Type III RM enzymes. The triplet motif GDN was found to be involved in interaction with the adenine of the methyl group donor AdoMet, while the other motifs were found to be taking part in Mod_A-Mod_B dimerization. A summary of the newly identified motifs of Group I of MTases of Type III RM enzymes is given in Table 4.2.

Table 4.1: N-terminal residues of EcoP15I involved in dimerization of Mod_A-Mod_B*

Mod A	*Motif*	*Mod B*	*Motif*	Distance (Å)
D30, Oδ2		S52, Oγ	Not conserved	2.9
Y56, ηOH		Q206, Nε2		3
S57, O		N59, N		3
E54, Oε1		G62, N		3.2
S55, O	Not conserved	N59, Nδ2		3.4
S172, Oγ		E54, Oε1		3.4
L58, O		K63, Nζ		3.5

*Residues from Mod_A and Mod_B are shown in green and cyan respectively. The height of letter represents the degree of conservation. Residues involved in Mod_A-Mod_B dimerization are highlighted by red dot.

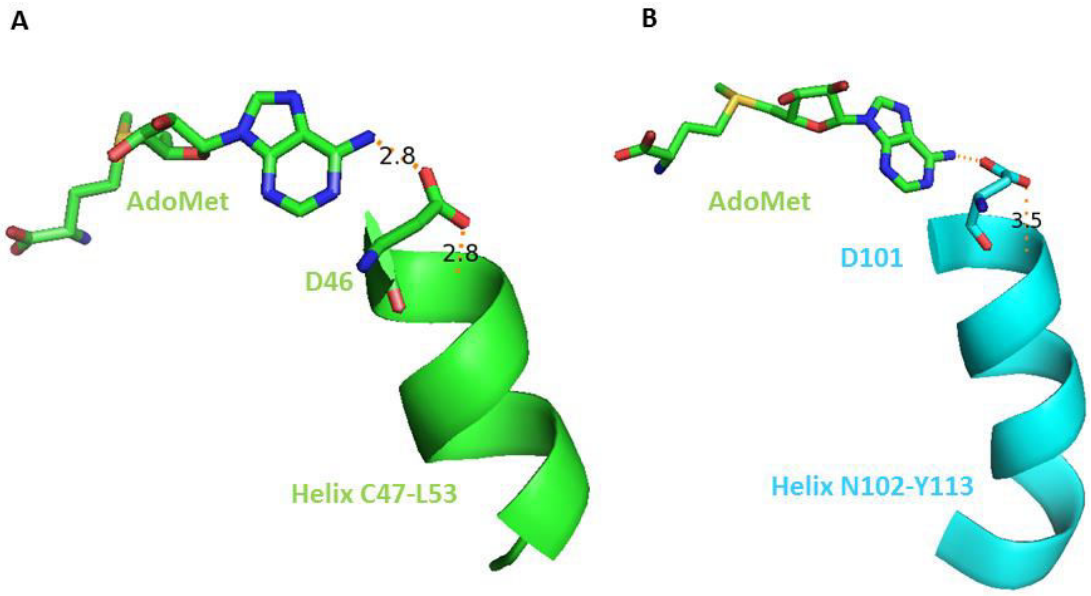


Figure 4.3: Role of aspartate in ß MTases. A] D46 (green sticks) interacts with N6 atom of AdoMet (green sticks) and caps the helix C47-L53 (green cartoon); B] D101 (cyan sticks) interacts with N6 atom of AdoMet (green sticks) and caps the helix N102-Y113 (cyan cartoon) AdoMet was modeled from M.RsrI (PDB ID: 1NW5) (green sticks). The distance between various atoms is shown as dashed lines. The numbers indicate distance in Å.

Table 4.2: Newly identified motifs of Group I of MTases of Type III RM enzymes*

Signature	EcoP15I coordinates	Function
	L21 to F35	Dimerization
	E54 to G62	Dimerization
	K63 to A70	Dimerization
	G100 to N102	Catalysis/ AdoMet Binding
	S168 to A173	Dimerization
	S205 to L207	Dimerization

*Equivalent motifs are present in Group II and III

3.1.2 N-terminal insertion in Mod of Type III RM enzymes

The structural superposition of M1.MboII with MTase of EcoP15I showed that there are three major insertions in EcoP15I. The first insertion is at N terminus. The second insertion results in a TRD larger in EcoP15 than in M1.MboII. The third insertion is at C terminus of Mod. Along with structural superposition, we also did a structure based sequence alignment of M1.MboII (PDB ID: 1G60), M.RsrI (PDB ID: 1NW5) and Mod_A of EcoP15I (PDB ID: 4ZCF) (Figure 4.4A). It was observed that the N-terminal insertion in EcoP15I is ~74 residues long, the TRD is ~133 residues long and the C-terminal insertion is ~132 residues long. To see if these insertions are conserved in Type III MTases, we calculated evolutionary conservation scores for amino acid positions in the MSA using ConSurf (18). These scores were then mapped on to the structure of Mod_A of EcoP15I (Figure 4.4C). It was observed that the N-terminal insertion is most conserved. The newly identified non-canonical motifs were located on the N-terminal insertion (Figures 4.2). In conclusion, N-terminal insertion is unique to Type III RM enzymes and contains conserved motifs involved in dimerization of the two Mods.

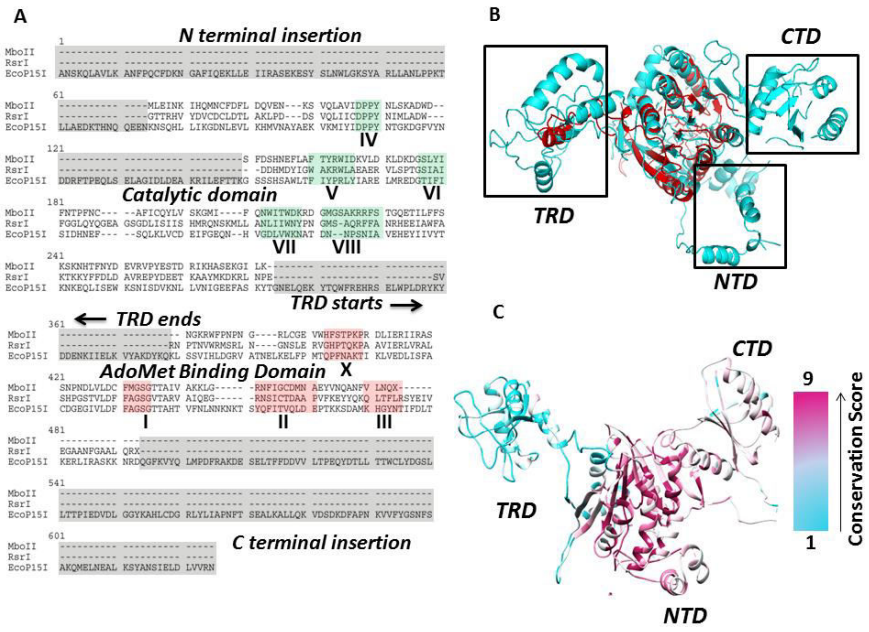


Figure 4.4: Insertions in MTase of Type III RM enzymes. A] Structure-based sequence alignment of M1.MboII, M.RsrI and EcoP15IMod_A. The figure shows NTD, catalytic domain, TRD and CTD. The insertion in EcoP15IMod_A is highlighted in grey. Canonical motifs are highlighted in green. B] Structural superposition of M1.MboII (PDB ID: 1G60) (in red) and EcoP15IMod_A (PDB ID: 4ZCF) (in cyan). Major insertions in M.EcoP15I are highlighted by rectangles; C] ConSurf conservation scores mapped on EcoP15IMod_A. The degree of conservation is depicted as heat map on right side, where cyan to pink transition represents lowest (1) to highest (9) ConSurf grade.

3.1.3 Analysis of lengths of various functional domains within Mod

The structure-based sequence analysis suggested that the MTase of Type III RM enzymes is composed of 5 different domains - an N-terminal domain (NTD), catalytic domain, TRD, AdoMet Binding domain and a C-terminal domain. Upon analyzing 1061 sequences of MTases of Type III RM enzymes, downloaded from REBASE on 12-9-2012, it was observed that the length of MTases showed large variations - ranging from as low as 428 to as long as 1357 residues (Figure 4.5A). To examine the cause of variation in length of MTases, we analyzed lengths of all 5 domains that constitute a full length MTase (Figure 4.5 B-F). This was done by plotting the length of each domain against their corresponding full length Mod.

The length of NTD showed variation ranging from as low as 17 residues to as long as 298 residues; however most of the members had length of NTD ranging from 50-100 residues (Figure 4.5B). We noticed that the lengths of catalytic domain (Figure 4.5C) and TRD (Figure 4.5D) were less variable. The length of AdoMet binding domain of Mod of length shorter than 1000 amino acids was less variable (Range: 60-120 residues). For Mods with length between 1000 and 1200, the length of AdoMet binding domain clustered between 200-250 residues; however for two Mods with a length greater than 1200, the length of AdoMet binding domain was greater than 500 residues. The most dramatic variation in length was observed for CTD where the length of CTD increased with an increase in length of Mod. The length of CTD showed three clusters, 50-150, 150-200 and 200-300 residues. This analysis revealed that the length variation in Mod originated mostly from variation in the length of CTD. We observed four prominent insertions in longer sequences that contributed to the increase in sequence length (Figure 4.6).

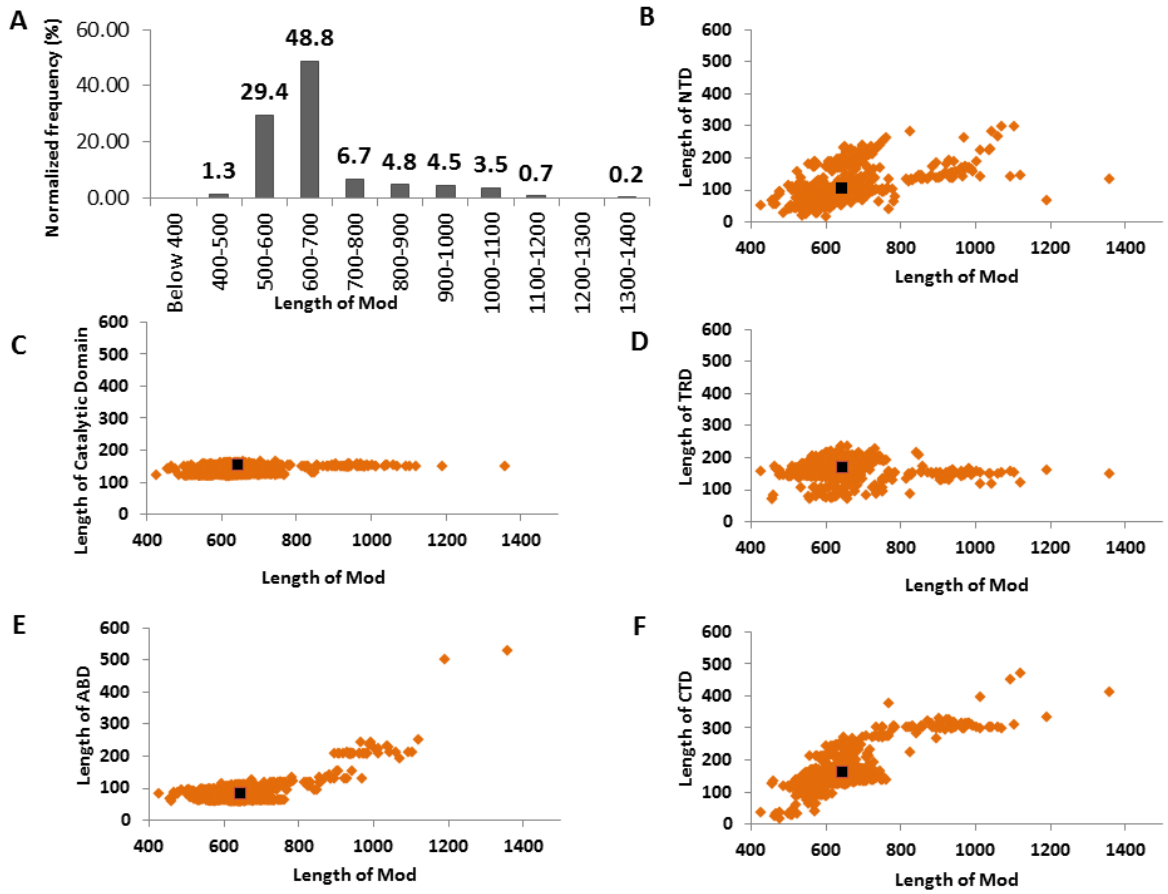


Figure 4.5: Analysis of lengths of Mod and various functional domains. A] Distribution of lengths of full length MTase represented as a bar graph, where X axis represents length in amino acids and Y axis represents normalized frequency of occurrence of a particular length range. Scatter plots of lengths of Mod (X axis) and lengths of N-terminal Domain [B], catalytic domain [C], TRD [D], AdoMet binding domain (ABD)[E] and CTD (Y axis). The location of EcoP1I is highlighted by black square in B-F. Total number of sequences = 1061.

To gain insights into the functional relevance of CTD we carried out a series of analyses. Firstly, evolutionary conservation scores were calculated using ConSurf (16), and buried surface area (BSA - solvent-accessible surface area of the corresponding residue that is buried upon interface formation) of residues of CTD using PDBePISA (29). The BSA score was calculated by assigning mnemonic score of 1 for 10% of the total solvent-accessible surface area buried. In the partial structure of EcoP15I co-crystallized with DNA substrate mimic and AMP, this C-terminal insertion of Mod_A is observed to be interacting with Res. Upon calculating the evolutionary conservation scores of residues comprising only CTD (residues 506-644, EcoP15I numbering), we observed that most of

the residues of CTD were positionally conserved. Out of 139 residues, 71 residues had a ConSurf grade greater than or equal to 6, while total 44 residues were observed to be taking part in interface between Mod_A and Res as identified by their BSA scores (Figure 4.6). Most of the residues at the interface were present at C-terminal part of CTD (Figure 4.6). From Res, a total of 74 residues were observed to take part in Mod_A-Res interface formation.

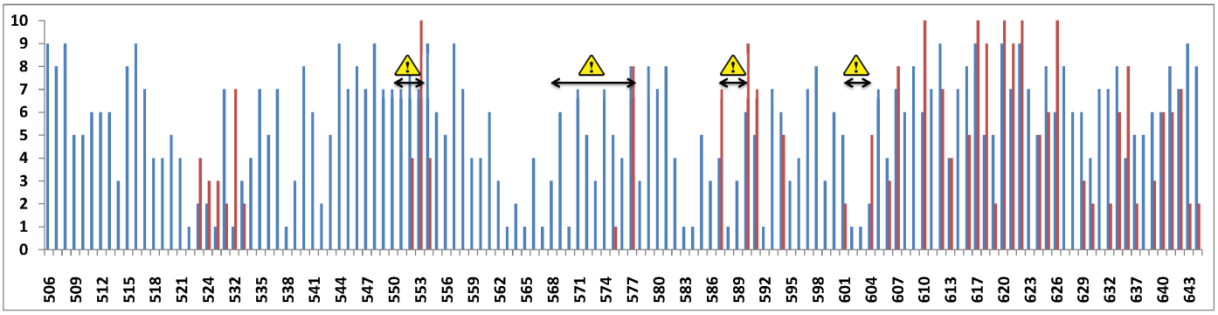


Figure 4.6: Positional conservation and interface scores of CTD. The X axis depicts residue positions in EcoP15I numbering, the Y axis depicts ConSurf grade and BSA scores on a scale of 0-10. Blue and maroon bars depict ConSurf conservation and BSA scores respectively. ⚠↔ depicts prominent insertions in the longer sequences.

3.2 Sequence analysis of ATPase domain

The Res subunit of Type III RM enzymes is a fusion of N-terminal motor domain (ATPase) and C-terminal endonuclease domain (12) (Figure 4.7A). The ATPase domain belongs to SF2 helicases (8). The core of SF2 helicases is made of a fusion of two similar subdomains, which have the fold of the bacterial recombinase RecA (Figure 4.7B, D, and E). Hence, these subdomains are also referred to as RecA1 and RecA2. The characteristic motifs of the helicase core can be classified based on their functions (Figure 4.7C), i.e. 1) ATP binding and hydrolysis, 2) nucleic acid binding, and 3) coupling ATP hydrolysis to its specific function. Recently, the structure of EcoP15I also revealed the presence of an additional all beta domain between RecA1 and RecA2, called the Pin domain (Figure 4.7A, B) (5).

Based on sequence alignment of 39 sequences of Res subunits of Type III RM enzymes, McClelland and Szczelkun identified the core helicase motifs in this type of enzymes (13). Since the structure of Res was unavailable then, a few motifs were declared as

putative and their possible functions were speculated. Recent determination of the structure of EcoP15I led to the identification of motifs Ib, Ic and IIa (5). These motifs are involved in contacting DNA and are not conserved. A complete list of these motifs is given in Chapter 1. We had initiated the sequence analysis of ATPase domain as an extension to McClelland and Szczelkun's study, before the structure of EcP15I was published. We aligned 96 sequences of the ATPase domain of Res by delineating it from cognate endonuclease domain. The selection of sequences was done as described in materials and methods. The boundary of the ATPase domain was set beyond motif VI (the last motif in RecA2; see Figure 4.7A). This analysis led to redefining one of the ATPase motifs.

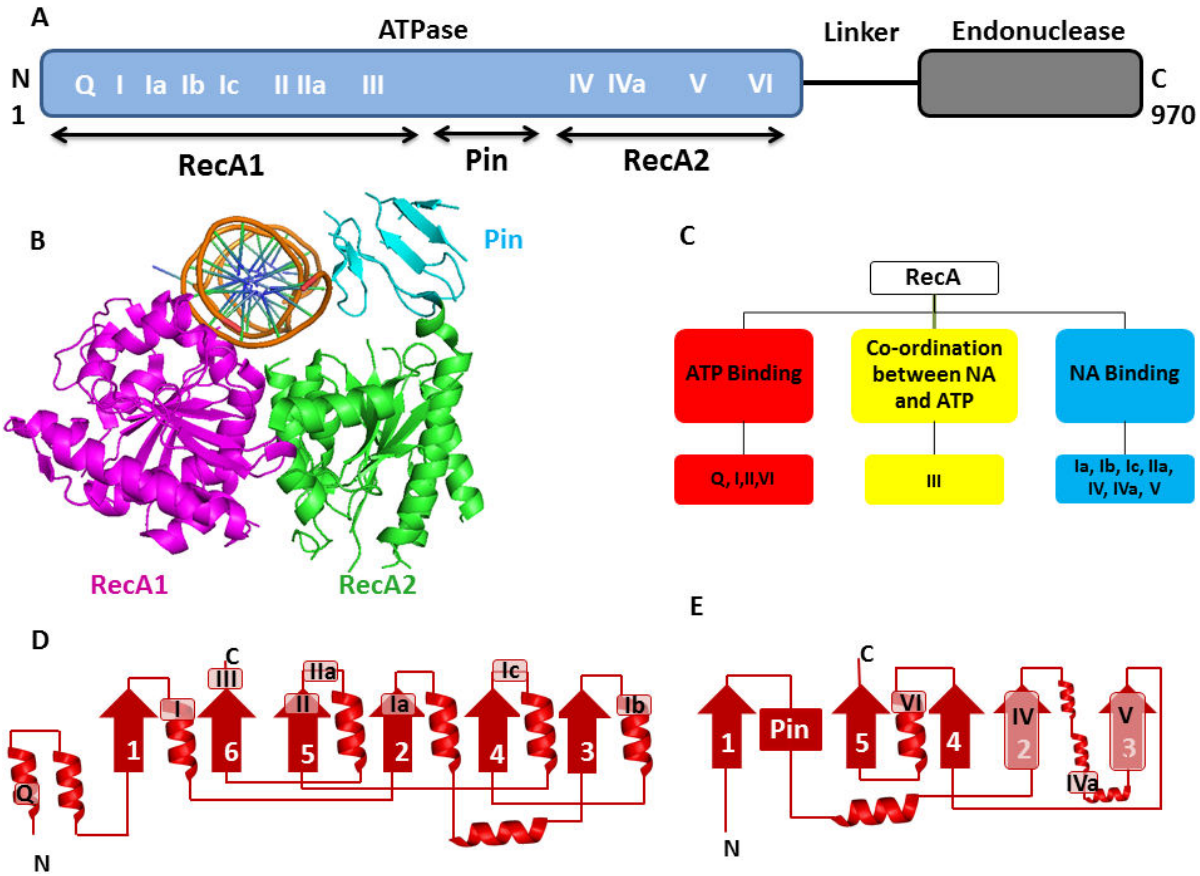


Figure 4.7: Organization of ATPase domain. A] Schematic of Res subunit of Type III RM enzymes showing fusion of ATPase and endonuclease domains; B] Cartoon representation of Res subunit shown as a part of the crystal structure of EcoP15I bound to DNA substrate mimic. The RecA1, RecA2 and Pin domain are shown in magenta, green and cyan respectively; C] Classification of various motifs in the RecA of an ATPase based on their functions; D] Topology of RecA1 showing locations of various motifs; E] Topology of RecA2 showing locations of various motifs. Beta strands and alpha helices are depicted as arrows and spirals respectively.

3.2.1 Identification of a unique non-canonical motif

To begin with, we built a phylogenetic tree from the MSA of ATPase domains. Similar to MTase domain, we observed three groups in the phylogenetic tree (Figure 4.8). Interestingly, the members within each of the three groups were identical for both MTase and ATPase domains. We saw that there are subtle differences between groups in canonical motifs Q, II, III, IIIa and V (Figure 4.9A). These differences contributed to clustering of different members in the three groups. We identified a short conserved signature **(K/R)(E/A)GT** in the RecA1 domain of Type III ATPases ahead of motif Ib (Figure 4.9B). We mapped this signature on the structure of EcoP15I co-crystallized with AMP and a DNA substrate mimic (PDB ID: 4ZCF). This conserved patch is located in the ATP binding pocket at the boundary of RecA1 and RecA2 (Figure 4.9B, C). The conserved glycine is involved in interaction with 3'-OH of the ribose of ATP through main chain oxygen. The conserved threonine makes a hydrogen bond with one of the oxygens of α phosphate (Figure 4.10). Earlier, McClelland and Szczelkun had predicted a signature for motif Ia as GxxKF++VVPSxAIKEG$(x)_{9\text{-}11}$EHF. This whole signature was thought to be involved in contacting DNA. After mapping this putative motif on EcoP15I structure, we propose that there are two different motifs within this signature:

1) GxxKF++VVPS: Motif Ia, involved in contacting DNA
2) (K/R)EGT: **A new motif**, involved in contacting ATP

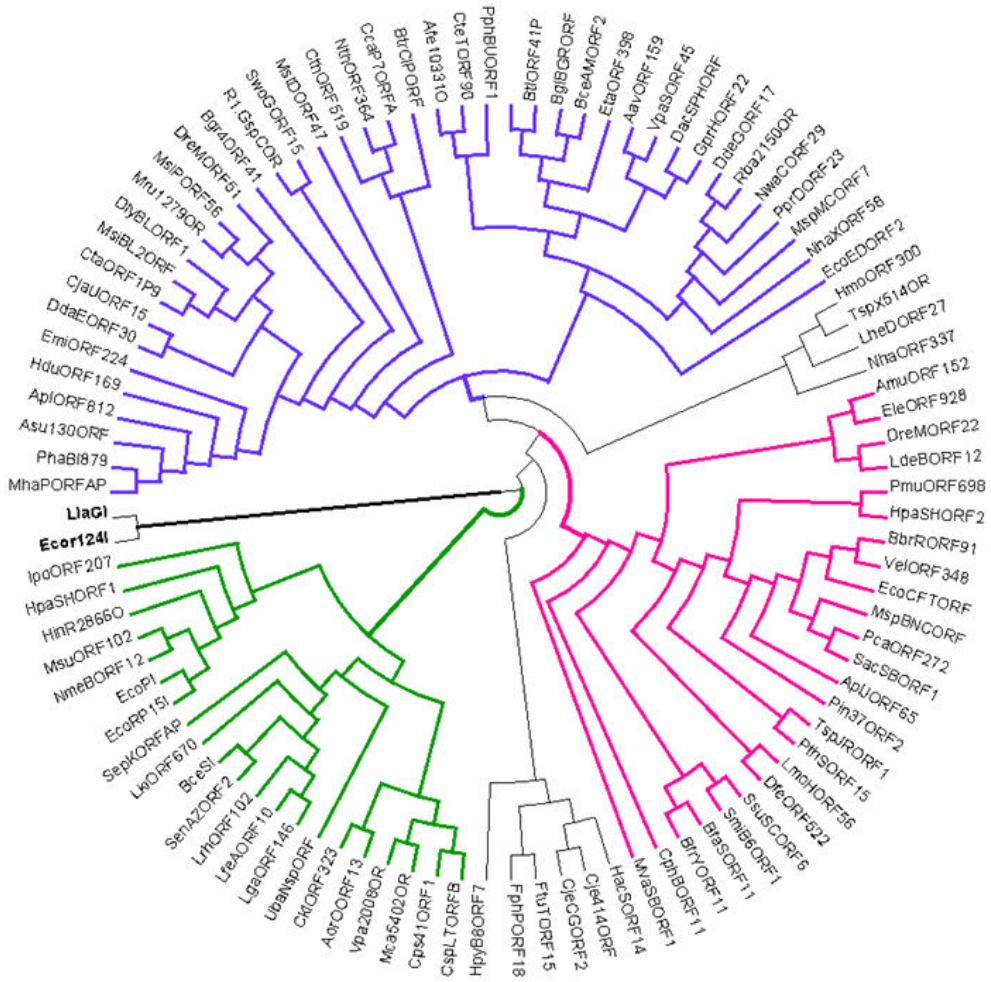


Figure 4.8: Phylogenetic analyses of ATPase associated with Type III RM enzyme. Rooted NJ phylogenetic tree of ATPase domain. ATPases of EcoR124I (Type I RM enzyme) and LlaGI (Type ISP RM enzyme) were used as outgroups shown as thick black line. The three groups are highlighted as green (Group I), pink (Group II) and blue (Group III). The outliers are shown in black lines.

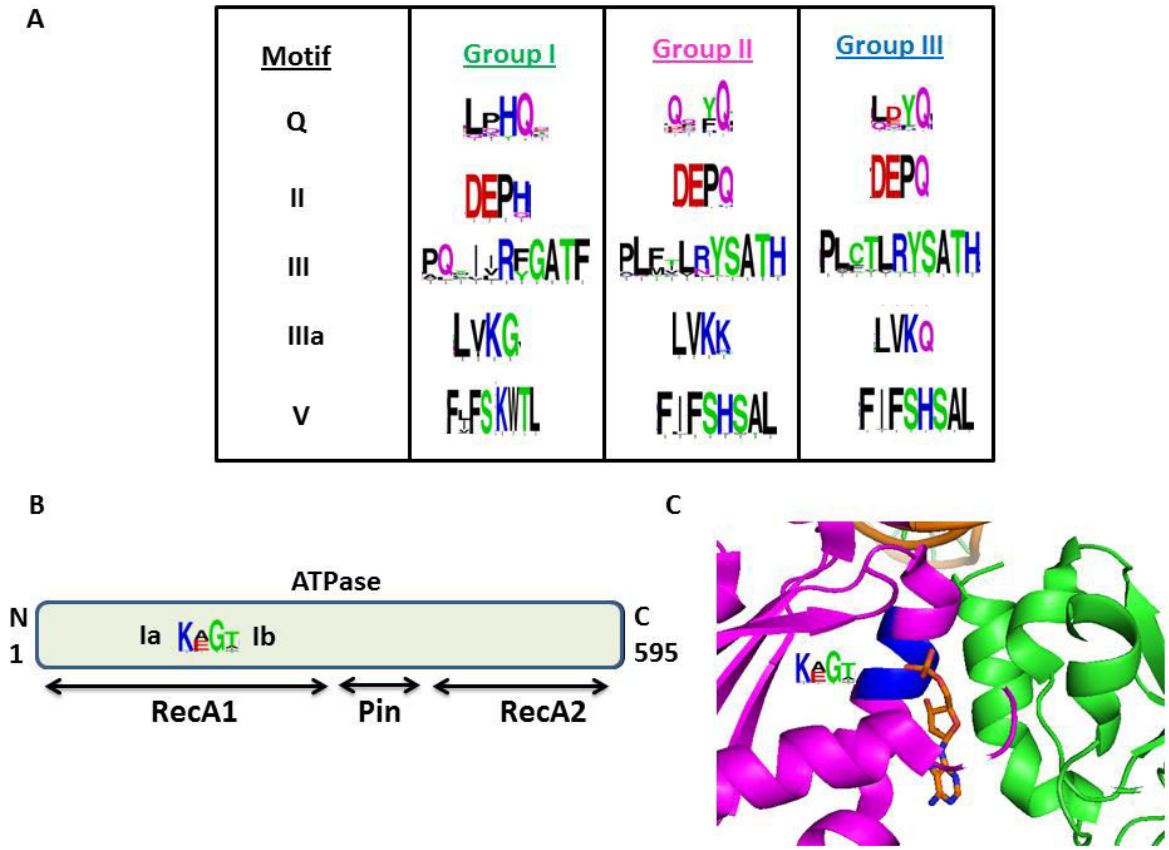


Figure 4.9: Sequence features of three phylogenetic groups in ATPase domain. A] Differences in canonical motifs of ATPase within groups. The height of letters denotes the degree of conservation. B] Location of conserved motif (K/R)(E/A)GT; C] Location of (K/R)(E/A)GT (shown as blue patch on RecA1) in three dimensional structure. AMP is shown as orange sticks.

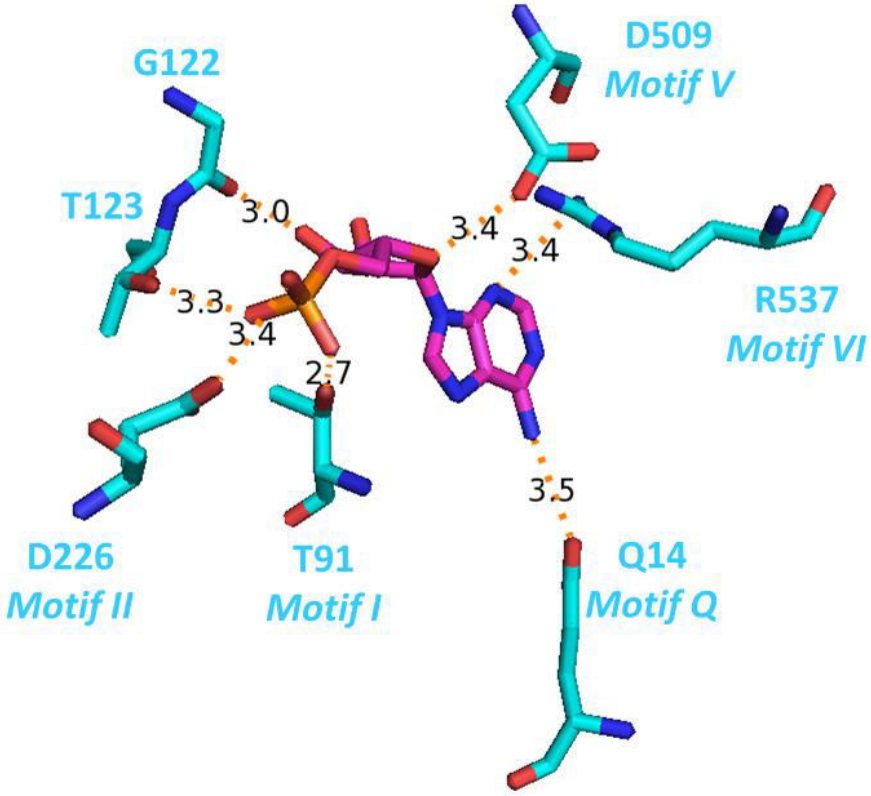


Figure 4.10: Interaction of G122 and T123 with AMP in EcoP15I (PDB ID: 4ZCF). AMP is shown as magenta sticks. The key residues interacting with AMP are shown as cyan sticks. The dashed lines denote distances between atoms. The numbers denote distance in Å.

3.2.2 Identification of Pin domain

Along with the highly conserved canonical and non-canonical motifs, we detected a weakly conserved patch between the last motif of RecA1 (Motif VI) and first motif of RecA2 (Motif IV) (Figure 4.11A and B). The average length of this patch was ~ 138 residues. Interestingly, on mapping this region onto the recently determined structure of EcoP15I, we found this region to be an insertion, which is called the Pin domain (5). From sequence alignment, we saw this domain to be present in all Type III RM enzymes, which indicates its functional importance. Based on our analyses, we suggest that the Pin domain is a unique and important feature of Type III RM enzymes. Despite the occurrence of this insertion in all the sequences analyzed, we could not find any conserved motifs in them.

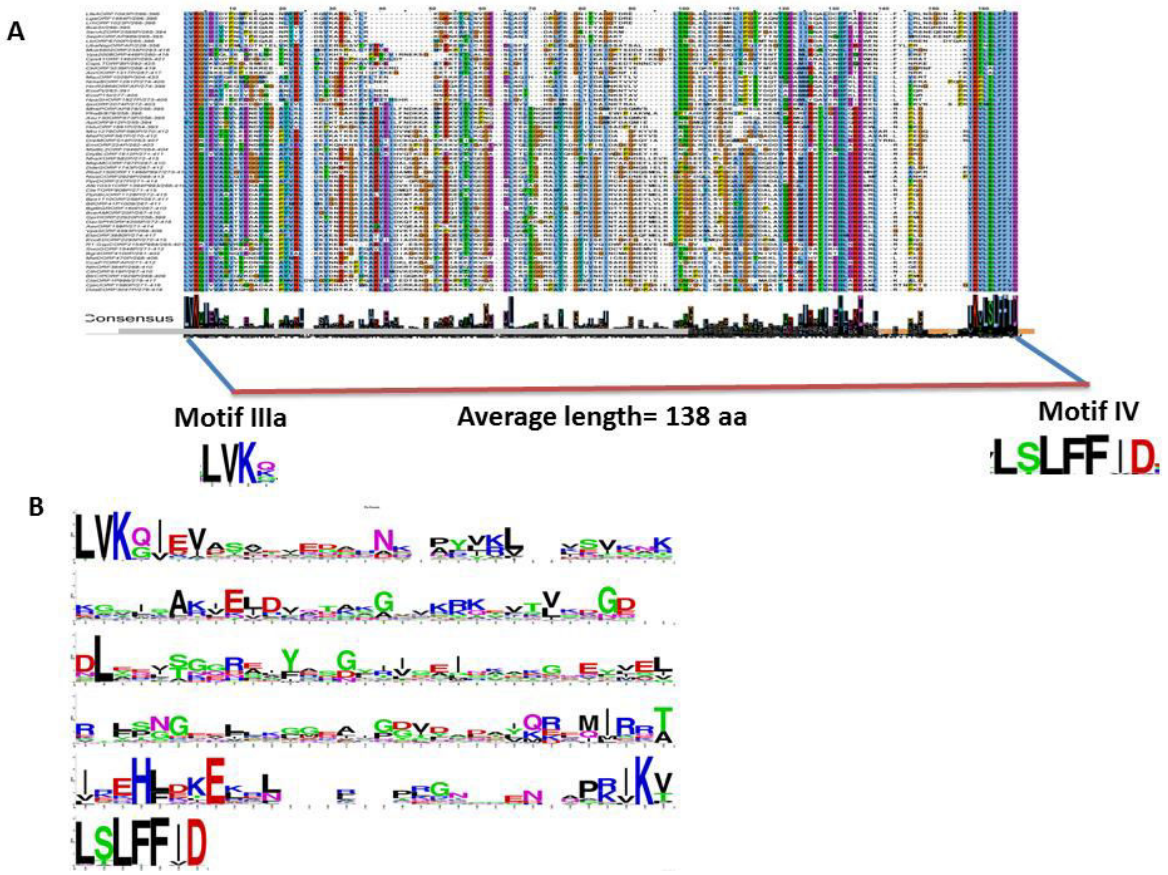


Figure 4.11: Analysis of Pin domain of Type III RM enzymes. A] MSA of ATPases associated with Type III RM enzymes showing the presence of Pin domain between motifs IIIa and IV; B] Alignment features of Pin domain. The height of letters depicts degree of conservation.

3.3 Analysis of endonuclease domain

The C-terminal endonuclease domain associated with Type III RM enzymes has motifs characteristic of Archeal Holliday Junction Resolvase (AHJR) family of nucleases (9). To examine the sequence features of this domain, we aligned 96 sequences of the nuclease domain of the Res subunit. A phylogenetic analysis of the Type III endonuclease domain using the MSA as input was carried out. The endonuclease domain of EcoR124I (Type I RM enzyme) which belongs to the RecB class of nucleases and LlaGI (Type ISP RM enzyme) which belongs to the Mrr class of nucleases were chosen as outgroups (1, 25). Once again we got three groups from the phylogenetic tree (Figure 4.12). Interestingly, the members of these three different groups were identical as were found in MTase and ATPase domains.

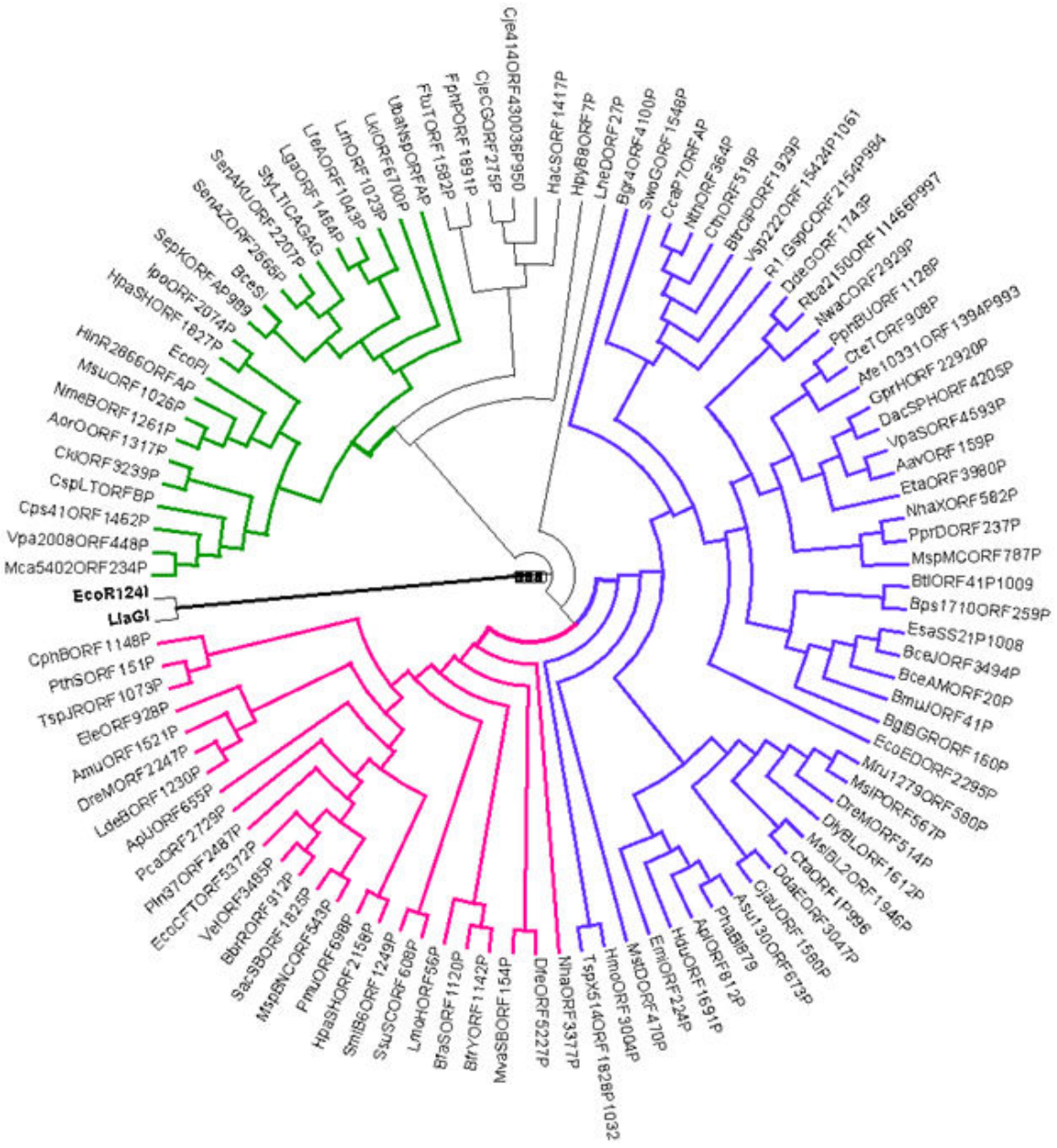


Figure 4.12: Phylogenetic analyses of endonuclease domain. Rooted NJ tree of endonuclease domain associated with Type III RM enzymes. Nucleases of EcoR124I (Type I RM enzyme) and LlaGI (Type ISP RM enzyme) were used as an outgroup and is shown by thick black line. The three groups are shown as green (group I), pink (group II) and blue (group III). The outliers are shown in black lines.

Although endonuclease domain of Type III RM enzymes belongs to the AHJR family of nuclease, we noticed that the length of endonuclease domain of Type III RM enzymes (~400) was much larger than that of a typical AHJR endonuclease (~140). Using limited proteolysis, mass spectrometry and insertional mutagenesis, it was previously reported that the ATPase and endonuclease domains are connected by a linker region of 23

residues (12). This linker region was a part of endonuclease in all our analyses. However, a linker length of 23 residues could not account for the observed large lengths of sequences of endonucleases in our analysis. We suspected that the linker between ATPase and endonuclease could be much longer than that predicted by limited proteolysis studies. This speculation was substantiated by the recently published partial structure of EcoP15I which lacked the endonuclease domain and a part of linker due to weak electron density (5). Figure 4.13 shows the missing regions of the Res subunit of EcoP15I. The motif VI, which is the last functional motif of the ATPase domain, spans from Q530 to L538 (QEVGLRGRL). After this motif, there are large numbers of missing residues along with full endonuclease domain, making the delineation of boundaries difficult.

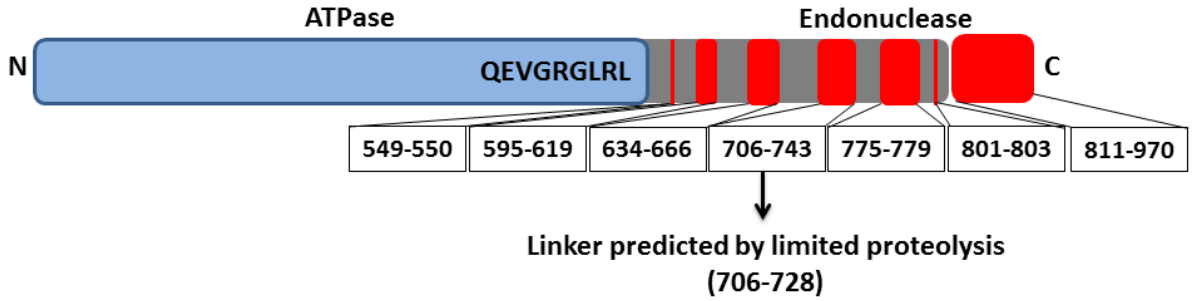


Figure 4.13: Missing regions of the Res subunit of EcoP15I (PDB ID: 4ZCF). Putative boundary of ATPase domain (blue rectangle) is assumed at motif VI (QEVGRGLRL). The missing regions in endonuclease domain (grey rectangle) are shown as red patches. Region spanning residues 706-728 is the linker predicted by limited proteolysis (12).

3.3.1 Delineation of boundary between linker and endonuclease

To predict the start point of the endonuclease domain, we used a combination of sequence-structure analysis. As the endonucleases associated with Type III RM enzymes belong to AHJR family of nucleases, we aligned 100 sequences of Holliday junction resolvases (Hjc) to see the conserved signatures (Figure 4.14). The members of this family are characterized by a set of 3 conserved motifs - I, II and III forming a metal ion binding pocket (2, 26–28).

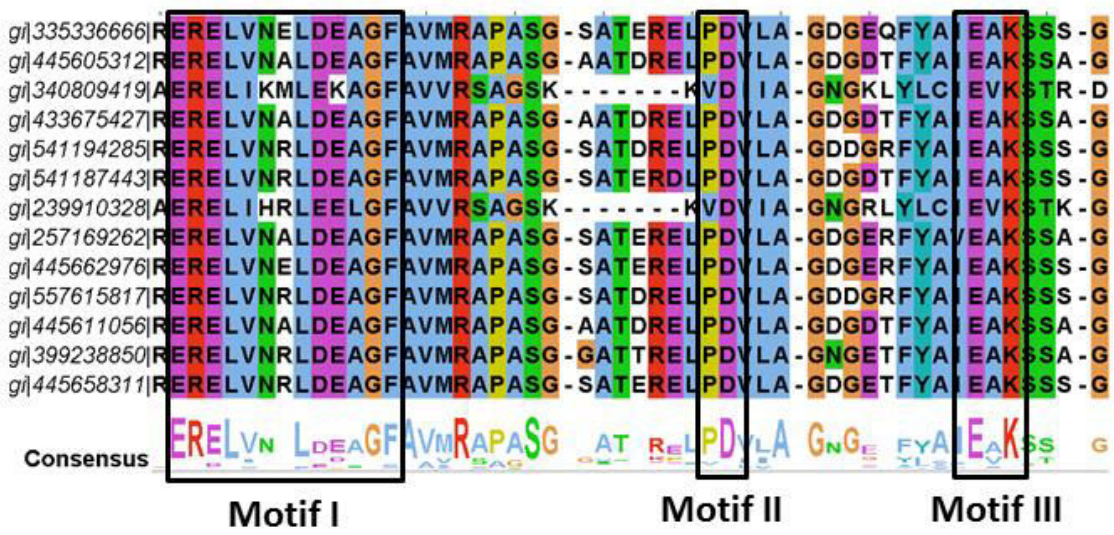


Figure 4.14: Sequence analyses of AHJR. A part of MSA of AHJR is represented to highlight canonical motifs I, II and III. The consensus sequence logo is shown at the bottom of the alignment where the height of letters represents the degree of conservation. Key canonical motifs are highlighted by black rectangle.

To delineate the boundary between N-terminal ATPase and C-terminal endonuclease, it was necessary to get insights on these canonical motifs. From the MSA of endonucleases of Type III RM enzymes, we could confidently assign motifs II (PD) and III (ExK) (Figure 4.15A). However assignment of motif I was not straightforward, since analogous conserved motif could not be detected in the MSA of Type III RM enzymes. Identification of motif I was important to assign a putative boundary between the linker and the endonuclease. Towards this, we did structure based sequence alignment of EcoP1I with five structures of Holliday junction resolvases. (PDB IDs: 1GEF, 1OB8, 4TKD, 2FCO and 2WCW)

Two important points were considered while assigning Motif I:

1) Is the signature conserved in Type III RM enzymes?
2) Is the chemical nature of conserved signature from Type III RM enzymes and canonical motif I of AHJRs similar?

Figure 4.15A shows putative motif I of endonuclease associated with Type III RM enzymes satisfying these criteria. The identification of putative motif I let us draw a boundary between ATPase and endonuclease domains.

Based on the results from structure based sequence analysis and partial structure of EcoP15I, we propose delineation of the linker between ATPase and endonuclease domains of EcoP1I/P15I as shown in Figure 4.15B. The delineation of boundaries between ATPase domain, linker and endonuclease domain indicated that the additional length of endonuclease domain observed was contributed by a large linker between ATPase and endonuclease domains. The linker comprises around 260 amino acids in EcoP1I/P15I. The C-terminal endonuclease is a short (~115 amino acids) domain.

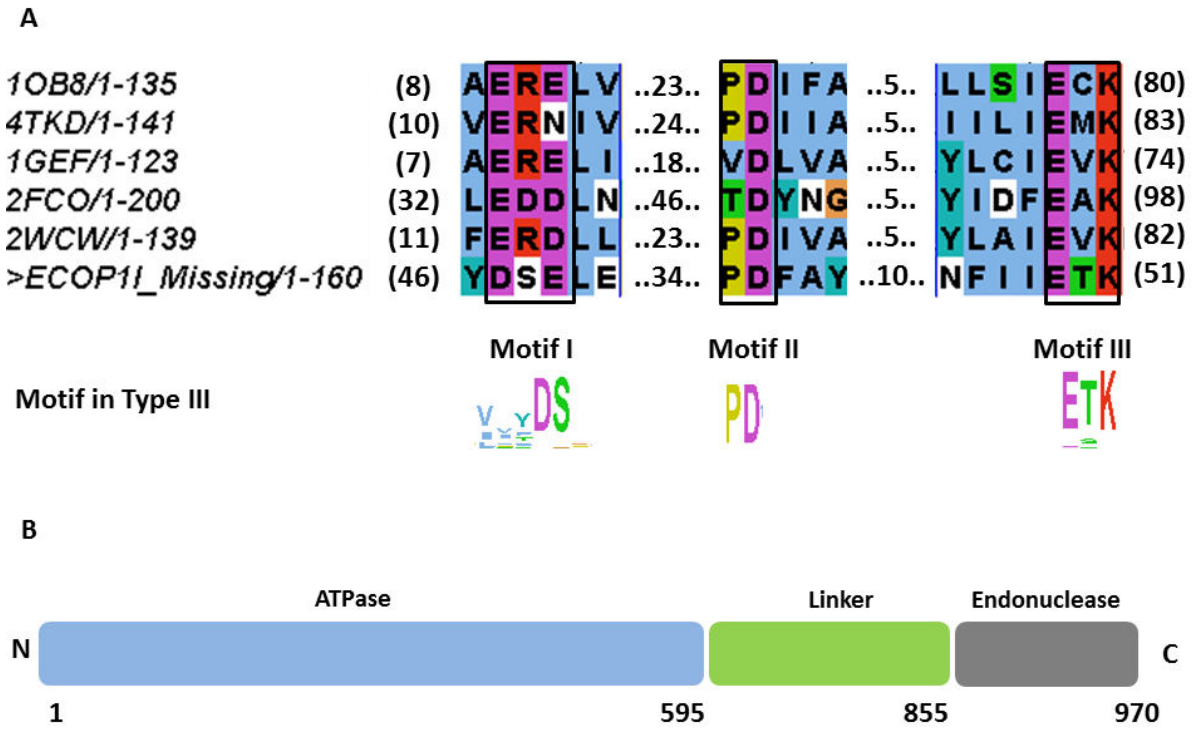


Figure 4.15: Identification of Motif I in endonuclease domain of Type III RM enzymes. Structure based sequence alignment of EcoP1I with five Hjc members depicting putative motif I of endonuclease of Type III RM enzymes. Key canonical motifs are highlighted by black rectangle. D] Approximate domain boundaries in Res subunit of EcoP1I/P15I. The linker (green) is inserted between ATPase (blue) and endonuclease (grey) domains.

4. Discussion

As a first step towards understanding the complexity of Type III RM enzymes, we took up sequence analysis of various domains within each subunit, i.e. Mod and Res. This helped us understand the canonical motifs in detail. An exhaustive sequence analysis of Type III RM enzymes led us to identify certain unique features at the sequence level. Structural and functional significance to these features could be attributed with greater confidence upon the availability of first structure of Type III RM enzyme, that of EcoP15I published in 2015 (5). When the coordinates of the structure became available, we revisited the sequence analysis to gain insights into the observations that we had made.

Methyltransferase: Type III RM enzymes have a cognate MTase, which catalyzes the transfer of a methyl group from S Adenosine methionine (AdoMet) to either adenine or cytosine base of DNA. An exhaustive sequence alignment of 92 sequences followed by a phylogenetic analysis resulted in 3 distinct clusters of sequences. We found that there were differential conserved patches of residues amongst these groups. Interestingly, apart from canonical motifs, only one non-canonical signature (GDN) was detected in all groups. The functional relevance of these non-canonical motifs could not be interpreted until Gupta et al published the partial structure of EcoP15I bound to a DNA substrate mimic (5). We mapped the motifs on the structure. We observed that six differentially conserved patches were all located at the N-terminus of Mod and these residues participated in Mod_A-Mod_B dimerization. The aspartate (D101 in EcoP15I) of non-canonical motif GDN was found to be involved in interaction with AdoMet and also capped the helix N102 to Y113. Since these are *in silico* predictions, we need to validate functional relevance of these motifs by in vitro mutational analysis.

We also compared the structure of EcoP15I MTase with known structures of two other ß class but stand-alone MTases, namely M1.MboII and M.RsrI. A combination of sequence-structure analyses demonstrated that there are three major insertions in Type III MTases, an N-terminal insertion, TRD and a C-terminal insertion. An analysis of 1061 sequences of MTases of Type III RM enzymes revealed that the length of Mod ranged from 428 to 1357 residues. We carried out a comparative study of lengths of five functional domains within MTases of Type III RM enzymes i.e. NTD, catalytic domain,

TRD, AdoMet binding domain and CTD. We noticed that the length of CTD increased corresponding to an increase in Mod length.

The evolutionary conservation scores and BSA scores of CTD calculated using coordinates of EcoP15I (PDB ID: 4ZCF) suggested that a significant portion of CTD is positionally conserved and is involved in making interface with Res. Mod-Res crosstalk is probably crucial for relaying multiple messages across domains, such as recognition of unmodified target site triggers ATPase activity of Res and hydrolysis of ATP, which in turn brings about large conformational change in Mod_2Res assembly to switch from loading to sliding mode. Gupta et al suggested that upon ATP hydrolysis, TRD_A would go away from target site to approach Pin such that enzyme remodels from a specific loading mode to long range sliding mode (5). We then sought to understand the requirement of CTD for interface formation with cognate Res. We observed that 16 Type III Mods having a CTD 64 residues long, did not have cognate Res in the ORF (Open Reading Frame). Only one Mod having a shorter CTD had a cognate Res. This suggested the importance of CTD for complex formation with Res. Those Mods having shorter CTD may act as orphan MTases (MTases not associated with cognate Res). We propose that the length of CTD of a Mod can be used to predict if it would function as an orphan MTase or be part of a canonical Type III RM enzyme assembly. Orphan MTases have been shown to have other cellular functions such as implications on epigenetic regulations by providing site-specific methylation (30) or providing immunity against parasitism by RM enzymes by targeting the same DNA sequence as the corresponding RM enzyme for example Dcm, a cytosine methyltransferase in *E. coli* that targets the same recognition site as the EcoRII RM enzyme (31, 32).

ATPase: The ATPase domain of Type III RM enzymes is located at the N terminus of Res. It belongs to SF2 helicases. For sequence analyses, based on the information on classical SF2 helicase motifs present in ATPase domain of Type III RM enzymes, we marked the boundary of ATPase domain at motif VI. MSA of 96 such sequences led to identification of an additional motif in RecA1 located at the junction of RecA1 and RecA2. This motif was considered as part of putative motif Ia ($GxxKF++VVPSxAIKEG(x)_{9\text{-}11}EHF$) involved in interacting with DNA (13). However, through a systematic sequence-structure analysis we redefined the role of this motif. A conserved threonine (T123 in EcoP15I)

interacts with alpha phosphate and a conserved glycine (G122 in EcoP15I) interacts with 3'OH of ribose of AMP (5). These two residues are part of a conserved motif (K/R)(E/A)GT in Type III RM enzymes. Functional relevance of this putative motif needs to be validated through mutagenesis studies. In structures of other SF2 helicases, despite the similarity in nucleotide binding pocket, nucleotide is held in the pocket via interaction with other motifs. For example, in the structure of NS3h (DENV) co-crystallized with AMPPNP and ssRNA, the N418 of motif V interacts with alpha phosphate of AMPPNP, while in the structure of VASA co-crystallized with AMPPNP and ssRNA, the interaction is accomplished by T296 of motif I (33).

Apart from identification of a new motif, we demonstrated that an insertion in RecA1 and RecA2, called the Pin domain, is present in all the members of Type III RM enzymes. Recently published partial crystal structure of EcoP15I revealed proximity of TRD_A to Pin domain (5). The DNA is bent by ~24° at the Mod_A-Res junction. Based on these observations, Gupta et al suggested that enzyme sliding is facilitated by interaction between TRD_A and Pin domain upon hydrolysis of ATP. The presence of Pin domain in all Type III RM enzymes corroborates the functional relevance of Pin domain in facilitating domain movements by acting as a hinge between Mod_A and Res. Though the Pin domain was found in all the Type III RM enzymes used in analysis, we did not find any conserved motif in Pin domain. Translocation assays using Pin-deleted constructs of EcoP1I or EcoP15I would give insights into requirement of this domain for large conformational changes accompanied with ATP hydrolysis.

Endonuclease: Endonuclease domain is located at the C terminus of Res. Detailed sequence/structure analysis of this domain is under-represented, apart from assigning the endonuclease domain to AHJR family of nuclease (9). Current structure of EcoP15I does not give any information about this domain due to weak electron density in this region. To understand sequence features of endonuclease domain of Type III RM enzymes, we compared endonuclease Type III RM enzymes with AHJR nucleases. To our surprise, the sequences of endonuclease domain of Type III RM enzymes were much larger. This indicated that the input sequences of Type III RM enzymes had an additional segment that was absent in the AHJR nucleases. According to limited proteolysis studies, the ATPase and endonuclease domains are connected by a 23 residue long

linker (12). We suspected that this linker could be much longer, based on length differences between endonuclease of Type III RM enzymes and AHJR. To delineate the boundaries between linker and endonuclease core, it was necessary to identify motif I (the first motif in the core). In 2000, Aravind et al (9) did a comprehensive sequence analysis of Holliday junction resolvases (Hjc) and proposed that the endonuclease fold is present in a variety of nucleases including Type II RM enzymes. This structural fold is a common feature of a large number of bacterial and viral nucleases. Based on the PSI-BLAST analysis of related nucleases, the authors classified the endonuclease fold into 2 groups:

1) Superfamily I: AHJR-Mrr family, RecB family, PHAC family and λ exonuclease family
2) Superfamily II: Vsr homologoues, classical restriction endonucleases

We took up structure-guided sequence alignment to decipher sequence characteristics of endonuclease domain of Type III RM enzymes. With an informed knowledge of key motifs of AHJR family of nucleases, we compared crystal structures of five AHJR nucleases. Motifs II and III could be assigned from MSA of endonucleases of Type III; however identification of motif I was not straightforward. With a structure-based sequence alignment, we proposed putative motif I (DS) (D858 and S859 in EcoP1I) for endonuclease associated with Type III RM enzymes. Motif I of the AHJR fold is characterized by a strongly conserved acidic residue (D/E). The acidic side chain in motif I has implications in stabilizing metal ion binding (34) or facilitating conformational transitions for coordination of the catalytic triad (35). In LlaGI like proteins (Type ISP RM enzymes) belonging to Mrr class of nucleases, this role is played by a conserved glutamate (1).

The *in silico* prediction of motif I forms the platform for future studies of validating its functional relevance experimentally. As Motif I, II and III form a metal ion binding pocket, the Mg^{2+} dependent cleavage competency of endonuclease mutants could be examined, where motif I is altered. Such studies have been done with EcoP1I for motifs II and III, where the mutants (D898A and E916A) were deficient in cleavage (36).

Phylogenetic analyses of Type III RM enzymes resulted in grouping of sequences in three groups. The members within each group were identical in all the functional modules, i.e. MTase, ATPase and endonuclease (Tables A1, A2, A3, Appendix I). This observation suggested an interesting possibility of co-evolution of Mod and Res in Type III RM enzymes. A detailed investigation on features of these groups such as comparison of characteristics of the host bacteria is required to understand the rationale behind such clustering. Additionally we observed few sequences lying outside these three groups. A similar analysis with larger set of sequences could help reconcile possible relatedness of these outliers with other sequences.

To summarize, an exhaustive sequence of Type III RM enzymes has led us to identify unique signatures. The analysis still needs to be supported with experimental evidence to validate the functional relevance of these signatures.

5. References

1. Smith,R.M., Josephsen,J. and Szczelkun,M.D. (2009) An Mrr-family nuclease motif in the single polypeptide restriction – modification enzyme LlaGI. *Nucleic Acids Res.*, **37**, 7231–7238.

2. Kosinski,J., Feder,M. and Bujnicki,J.M. (2005) The PD- (D / E) XK superfamily revisited : identification of new members among proteins involved in DNA metabolism and functional predictions for domains of (hitherto) unknown function. *BMC Bioinformatics*, **6**, 1–13.

3. Malone,T., Blumenthal,R.M., Cheng,X. and Structural,W.M.K. (1995) Structure-guided analysis reveals nine sequence motifs conserved among DNA amino-methyl-transferases , and suggests a catalytic mechanism for these enzymes. *J. Mol. Biol.*, **253**, 618–632.

4. Aroul-Selvram,R., Hubbard,T. and Sasidharan,R. (2009) Domain insertions in protein structures. *J. Mol. Biol.*, **338**, 633–641.

5. Gupta,Y.K., Chan,S.-H., Xu,S. and Aggarwal,A.K. (2015) Structural basis of asymmetric DNA methylation and ATP-triggered long-range diffusion by EcoP15I. *Nat. Commun.*, **6**, 1–10.

6. Wyszomirski,K.H., Curth,U., Mackeldanz,P., Schutkowski,M., Kru,D.H. and Mo,E. (2011) Type III restriction endonuclease EcoP15I is a heterotrimeric complex containing one Res subunit with several DNA-binding regions and ATPase activity. *Nucleic Acids Res.*, 10.1093/nar/gkr1239.

7. Butterer,A., Pernstich,C., Smith,R.M., Sobott,F., Szczelkun,M.D. and Tóth,J. (2014) Type III restriction endonucleases are heterotrimeric: comprising one helicase-nuclease subunit and a dimeric methyltransferase that binds only one specific DNA. *Nucleic Acids Res.*, **42**, 5139–50.

8. Gorbalenya,A.E. and Koonin,E. V (1991) Endonuclease (R) subunits of type-I and type -III restriction-modication enzymes contain a helicase-like domain. *FEBS Lett.*, **291**, 277–281.

9. Aravind,L., Makarova,K.S. and Koonin,E. V (2000) Holliday junction resolvases and

related nucleases : identification of new families , phyletic distribution and evolutionary trajectories. *Nucleic Acids Res.*, **28**, 3417–3432.

10. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

11. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2015) REBASE- A database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, **43**, D298-9.

12. Wagenführ,K., Pieper,S., Mackeldanz,P., Linscheid,M., Krüger,D.H. and Reuter,M. (2007) Structural domains in the type III restriction endonuclease EcoP15I: characterization by limited proteolysis, mass spectrometry and insertional mutagenesis. *J. Mol. Biol.*, **366**, 93–102.

13. Mcclelland,S.E. and Szczelkun,M.D. (2004) The Type I and III Restriction Endonucleases : Structural Elements in Molecular Motors that Process DNA. **14**.

14. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R., *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

15. Gouy,M., Guindon,S. and Gascuel,O. (2010) SeaView Version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* , **27**, 221–224.

16. Waterhouse,A.M., Procter,J.B., Martin,D.M.A., Clamp,M. and Barton,G.J. (2009) Jalview Version 2- A multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.

17. Crooks,G.E., Hon,G., Chandonia,J.-M. and Brenner,S.E. (2004) WebLogo: A sequence logo generator. *Genome Res.*, **14**, 1188–1190.

18. Ashkenazy,H., Erez,E., Martz,E., Pupko,T. and Ben-Tal,N. (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.*, **38**, W529–W533.

19. Pei,J., Kim,B.-H. and Grishin,N. V (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.*, **36**, 2295–2300.

20. Felsenstein,J. (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.

21. Uson,D.A.H.H. and Cornavacca,C.E.S. (2012) Dendroscope 3 : An interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.*, **61**, 1061–1067.

22. Holm,L. and Park,J. (2000) DaliLite workbench for protein structure comparison. *Bioinforma.* , **16**, 566–567.

23. Schrodinger LLC (2015) The PyMOL molecular graphics system, Version 1.8.

24. Pettersen,E.F., Goddard,T.D., Huang,C.C., Couch,G.S., Greenblatt,D.M., Meng,E.C. and Ferrin,T.E. (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.

25. Stanley,L.K., Weiserova,M., Szczelkun,M.D. and S,E. (2008) A RecB-family nuclease motif in the Type I restriction endonuclease EcoR124I. *Nucleic Acids Res.*, **36**, 3939–3949.

26. Lilley,D.M.J. and White,M.F. (2001) The junction-resolving enzymes. *Nat. Rev.*, **2**, 433–443.

27. Sukackaite,R., Lagunavicius,A., Stankevicius,K., Siksnys,V. and Urbanke,C. (2007) Restriction endonuclease BpuJI specific for the 5' -CCCGT sequence is related to the archaeal Holliday junction resolvase family. *Nucleic Acids Res.*, **35**, 2377–2389.

28. Nishino,T., Komori,K., Tsuchiya,D., Ishino,Y. and Morikawa,K. (2001) Crystal structure of the archaeal holliday junction resolvase Hjc and implications for DNA recognition. *Structure*, **9**, 197–204.

29. Evgeny,K. and Kim,H. (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.*, **43**.

30. Peterson,S.N. and Reich,N.O. (2008) Competitive Lrp and Dam assembly at the pap regulatory region: implications for mechanisms of epigenetic regulation. *J. Mol. Biol.*, **383**, 92–105.

31. Takahashi,N., Naito,Y., Handa,N. and Kobayashi,I. (2002) A DNA methyltransferase can protect the genome from postdisturbance attack by a restriction-modification gene complex. *J. Bacteriol.*, **184**, 6100–6108.

32. Seshasayee,A.S.N., Singh,P. and Krishna,S. (2012) Context-dependent conservation of DNA methyltransferases in bacteria. *Nucleic Acids Res.*, **40**, 7066–7073.

33. Gu,M. and Rice,C.M. (2010) Three conformational snapshots of the hepatitis C virus NS3 helicase reveal a ratchet translocation mechanism. *PNAS*, **107**, 521–528.

34. Galburt,E.A. and Stoddard,B.L. (2002) Catalytic mechanisms of restriction and homing endonucleases. *Biochemistry*, **41**, 13851–13860.

35. Horton,N.C. and Perona,J.J. (2004) DNA cleavage by EcoRV endonuclease: two metal ions in three metal ion binding sites. *Biochemistry*, **43**, 6841–6857.

36. Janscak,P., Sandmeier,U., Szczelkun,M.D. and Bickle,T.A. (2001) Subunit assembly and mode of DNA cleavage of the Type III restriction endonucleases EcoP1I and EcoP15I. *J. Mol. Biol.*, **306**, 417–431.

CHAPTER 5

Crystallographic Studies of EcoP1I

Chapter 5

Crystallographic Studies of EcoP1I

1. Introduction

Type III RM enzymes form a 260 kDa trimeric assembly of a two MTases (Mod) and a restriction endonuclease (Res) (1–3). Each subunit within the assembly is a combination of various specialized domains working in concert to bring about endonucleolytic cleavage of DNA (4–7). Mod consists of N-terminal catalytic domain, TRD and a C-terminal AdoMet binding domain (5); whereas Res comprises N-terminal ATPase and a C-terminal endonuclease domain connected by a flexible linker (8, 6, 7). The ATPase domain is composed of two RecA-like sub-domains namely RecA1 and RecA2 (3, 9). Together, this multi-domained assembly carries out two counteracting activities - methylation of the target site for protection of self-DNA from destructive cleavage and nucleolytic cleavage of the unmodified foreign DNA (10, 11). The cleavage requires ATP-induced long range communication between two inversely oriented unmodified target sites (12, 13, 2, 3, 14). This is achieved by a large conformational change from a loading to sliding mode triggered by ATP hydrolysis (12, 14). Although, ATPase domain of Type III RM enzyme harbours motifs characteristic to classical ATPase motors, Type III RM enzymes consume as little as 30 ATP molecules for communicating over large distances as opposed to classical ATPases (12), where 1 ATP is used per base pair (15, 16). The long range communication results in collision of an ATP-activated diffusing enzyme with a target site bound enzyme (14). The endonuclease domains of two colliding enzymes interact to bring about dsDNA break close to the target bound enzyme (17).

Although ATP-induced long-lived DNA sliding is a looming theme in Type III RM enzymes, mechanistic details about this energy efficient process still remain unanswered. Recently a partial crystal structure of EcoP15I bound to a DNA substrate mimic in presence of AMP reported “division of labour” between two Mods, where one of the Mods is responsible for recognition of target site while the other participates in methylation of target adenine (3). The structure also revealed presence of an additional Pin domain within the ATPase domain of Res. The authors proposed a hinge-like role for Pin domain to facilitate enzyme remodelling from a tighter loading mode to a much slackened diffusive mode. This structure, however, lacks information about the

complete linker and the entire endonuclease domain due to weak electron densities. In the current structure, the RecA2 domain is steered away from DNA. Hence, the molecular interactions of DNA with RecA2, linker and endonuclease domain are still obscure. We yet do not know, what is the precise nature of ATP-induced conformational change leading to a much distinct sliding mode, how do DNA-Mod and DNA-Res contacts get reorganised after ATP hydrolysis and how does Mod slacken the clasp on DNA to leave the target site.

To gain structural insights into the mechanism of Type III RM enzymes by using the tools of X-ray crystallography, EcoP1I, a prototypical Type III RM enzyme was chosen as a candidate. The primary aim was to solve the crystal structure of DNA bound EcoP1I. We set up crystallization trails with a suitable DNA substrate mimic. The initial crystal hits were improved by optimization of conditions. After screening a large number of crystals, one of them diffracted to 4.4 Å. Although the diffraction data is anisotropic, we have been able to do molecular replacement using partial structure of EcoP15I as search model.

The details of crystallization trials and efforts towards structure determination are presented in this chapter.

2. Materials and Methods

2.1 DNA substrates

Oligomers were purchased from Integrated DNA Technologies, USA, and Sigma-Aldrich, USA. DNA duplexes were made by annealing the single-strands and further purified using an 8 ml MonoQ™ 10/100 GL column. The resulting duplexes were washed thoroughly with MilliQ-purified water to remove any salt and concentrated using Vivaspin®2 (GE Healthcare) concentrator (MWCO 3kDa). The concentrated oligomer was stored in MilliQ-purified water at -80°C until further use. A list of dsDNA used for the crystallization experiments are listed in Table 5.1. The T (top strand) and B (bottom strand) pairs were annealed to get a duplex DNA. The convention to label the DNA substrate is as described in section 2.4, Chapter 2.

Table 5.1: Oligomers used for crystallization of EcoP1I

Name	Sequence (5' -> 3')	Length
5/26_P1_T	GTACTAGACCTATCCTGTATGCTACGTATTCGTATC	36
5/26_P1_B	GATACGAATACGTAGCATACAGGATAGGTCTAGTAC	36
6/26_P1_T	CGTACTAGACCTATCCTGTATGCTACGTATTCGTATC	37
6/26_P1_B	GATACGAATACGTAGCATACAGGATAGGTCTAGTACG	37
7/26_P1_T	ACGTACTAGACCTATCCTGTATGCTACGTATTCGTATC	38
7/26_P1_B	GATACGAATACGTAGCATACAGGATAGGTCTAGTACGT	38

2.2 Protein purification

EcoP1I$^{N\text{-}His}$ and EcoP1I (E916A) $^{C\text{-}His}$ were purified as described in section 2.3, Chapter 2. Selenomethionine (Se-Met) derivative of EcoP1I$^{N\text{-}His}$ was purified by a protocol similar to that of EcoP1I$^{N\text{-}His}$ by maintaining 6 mM DTT in all buffers.

2.3 Crystallization and optimization of conditions

A complex of the purified protein and DNA substrate was formed by mixing the two in 1:1.3 molar ratio at 4°C in crystallization buffer (10 mM Tris-Cl, pH 8, 100 mM NaCl, 1 mM DTT). The final crystallization mix contained 1 mM sinefungin and a protein concentration of 10 mg/ml. The sample was centrifuged at 15,000 rpm for 10 minutes prior to setting up crystallization trails. A crystallization drop size of 200 nL was used for all the initial crystallization trials by sitting drop vapour diffusion method. The nanodrops were set using a robotic liquid handler. An array of 1404 conditions was set

up as an initial screen for crystallization. The initial hits obtained from the screen were further optimized by setting up an additive screen (Hampton Research) with a suitable parent condition (PC). Further, combinations of various additives were used to optimize the quality of crystals. Recipes of these crystallization cocktails are listed in Table 5.2. For setting up bigger drops by sitting drop vapour diffusion method, 48 well crystallization plates were used with 1:1 protein to reservoir ratio. The total size of the drop was 2 µL (1 µL DNA-Protein complex + 1 µL reservoir solution). The volume of reservoir was 100 µL.

Table 5.2: Recipes of crystallization cocktails

Sr No	Crystallization Cocktail	Abbreviation
1	PC + KCl+ DCM+ Werner's Reagent	M1
2	PC + KCl+ DCM+ LiCl	M2
3	PC + KCl+ DCM	M3
4	PC+ KCl+ DCM+ Werner's Reagent+ Phenol	M4
5	PC+ KCl+ DCM+ NaBr	M5
6	PC+ KCl+ DCM+ EDTA	M6
7	PC+ KCl+ DCM+ Werner's Reagent+ PEG 3350	M7
8	PC+ Sorbitol	M8
9	PC+ PEG 3350	M9
10	PC+ Werner's Reagent+ NDSB 201	M10
11	PC+ Werner's Reagent+ NDSB 195	M11

2.4 Synthesis of hexamine cobalt (III) chloride (Werner's reagent)

Hexamine cobalt (III) chloride was synthesized from cobalt chloride (Sigma) and ammonium chloride (Sigma) (Figure 5.1A). Cobalt (II) was oxidized to cobalt (III) in ammonia rich medium. The hexamine cobalt (III) ion was stabilized by high concentration of ammonium salt. The reaction of conversion of cobalt (II) to cobalt (III) is as follows:

$$4CoCl_2 \cdot 6H_2O + 4NH_4Cl + 20NH_3 + O_2 \rightarrow 4[Co(NH_3)_6]Cl_3 + 8H_2O$$

20 mM $CoCl_2.6H_2O$ was mixed with 20 mM NH_4Cl thoroughly with MilliQ-purified water (20 mL). The mixture was swirled to dissolve solids. 0.5 g activated charcoal was mixed with the mixture (Figure 5.1B). The activated charcoal contains finely divided carbon sheets having large specific surface area. This helps trap the ligand molecules for ligand exchange. To this, 15 mL ammonia solution was added. To facilitate oxidation, air was bubbled in the Erlenmeyer flask (Figure 5.1C). Once the solution became brown, oxidation was stopped (2 hours). Crystals and charcoal were filtered through a Buchner funnel and further washed with hydrochloric acid (HCl) solution (3 mL conc. HCl in 40 mL water) (Figure 5.1D). The mixture was heated on a hot plate and filtered hot. The solution was cooled to 0° (Figure 5.1E) and product was crystallized by slowly adding concentrated HCl. The crystals of hexamine cobalt (III) chloride were filtered through Whatman® Grade 1 filter paper. The crystals were further washed with 65% and 95% ethanol. The product was dried at 80° C in oven for 1 hour (Figure 5.1F). The purity of the compound was checked by UV-Vis spectroscopy. Figure 5.1G shows the UV-Vis spectra of the synthesized compound. The two peaks on the spectrum are d-d transitions for the d6 cobalt complex.

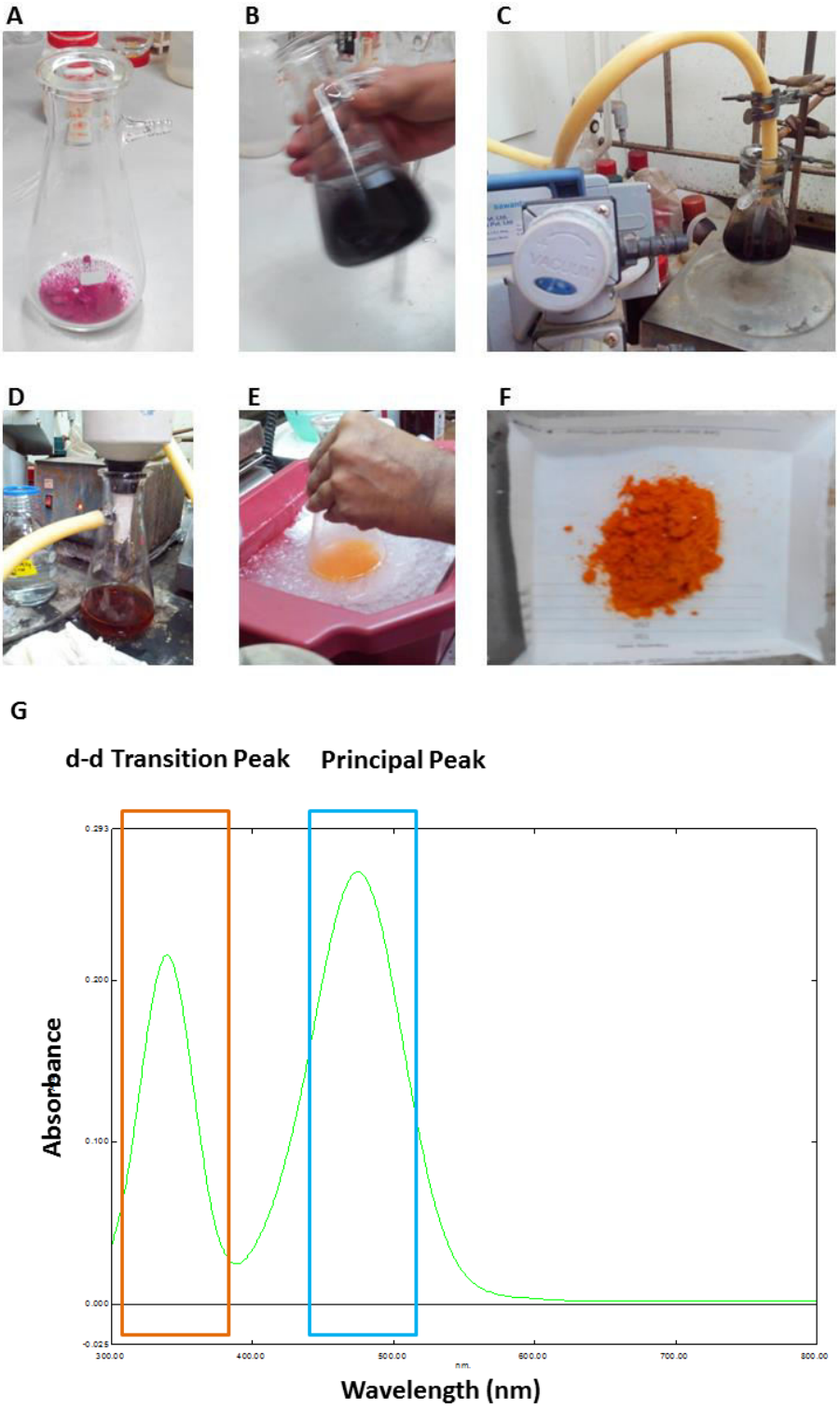


Figure 5.1: Synthesis of Hexamine cobalt (III) chloride. A-F] Step by step procedure of the synthesis protocol (see text above); G] UV-Vis spectra of Hexamine cobalt (III) chloride. X axis depicts wavelength in nm. Y axis depicts absorbance.

2.5 X-ray data collection and processing

Crystals that diffracted the best were cryo-protected for 1-2 S in reservoir buffer containing 35% glycerol. X-ray diffraction data were collected at the in-house diffraction facility (Rigaku MicroMax 007 X-ray generator and Mar Research 345D detector) and at synchrotron facilities at Diamond Light Source (DLS), Oxfordshire, UK and European Synchrotron Radiation Facility (ESRF). The diffraction data were indexed and processed using MOSFLM (18) or XDS (19). The processed data were scaled and merged using the program AIMLESS (20) in the CCP4 suite (21). 5% of the reflections were set aside for calculation of R_{free} during structure refinement.

2.6 Structure solution by molecular replacement

A structure solution for EcoP1I-DNA crystal was obtained by molecular replacement (MR) using the program PHASER (22). The coordinates of EcoP15I-DNA (PDB ID: 4ZCF) (3) served as initial search model. Coordinates of the DNA were not included in the model. The MR solution was refined against the data using the protocol of rigid body refinement implemented in the program PHENIX (23). The electron density map calculated using the partial structure was viewed using COOT (24).

3. Results

3.1 Initial optimization

In order to co-crystallize EcoP1I and its DNA substrate mimic, we made use of results from initial binding studies with EcoP1I as described in Chapter 2. These studies indicated that the purified EcoP1I (E916A) could bind DNA bearing single target site. The crystallization efforts were initiated by screening 1404 different conditions with EcoP1I (E916A)-SF-DNA (5/26_P1). We got 3 hits from the initial screen. The compositions of reservoir buffer of these 3 hits are given in Table 5.3. Upon comparing the quality of crystals in these three conditions, we observed that condition 2 had better quality crystals than either condition 1 or 3. To see the effect of pH on the quality of crystals we varied the pH of Na-HEPES buffer from 6.5 to 7.7. At pH less than or equal to 6.7 and greater than 7.5, only crystalline matter could be visualized, however conditions where pH values were maintained between 6.8 and 7.5, crystals could be seen. All further crystallization experiments were carried out using condition 2 as parent condition (PC) with pH 7.5.

Table 5.3: Composition of initial conditions that yielded crystals

	Condition 1	**Condition 2**	**Condition 3**
Precipitant Mix	PEG 4K, 10%, Glycerol, 20%	PEG 6K, 10%, MPD, 5%	PEG 8K, 10%, Ethylene glycol, 8%
BUFFER	MES-Imidazole, 0.1 M	Na-HEPES, 0.1 M	Na-HEPES, 0.1 M
LIGAND	Amino Acid Mix, 0.1 M	-	-
pH	6.5	7.5	7.5

To further improve quality of crystals, a variety of additives were added to parent condition 2 (Table 5.3). This was done by using the additive screen (Hampton Research). The crystals appeared in 4-5 days. After screening 96 additives, we got crystals in 21 conditions. The crystals resulting from additive screen were mostly clusters or micro-crystalline (Figure 5.2, A-K). However, with hexamine cobalt (III) chloride very tiny single crystals were observed (Figure 5.2, L).

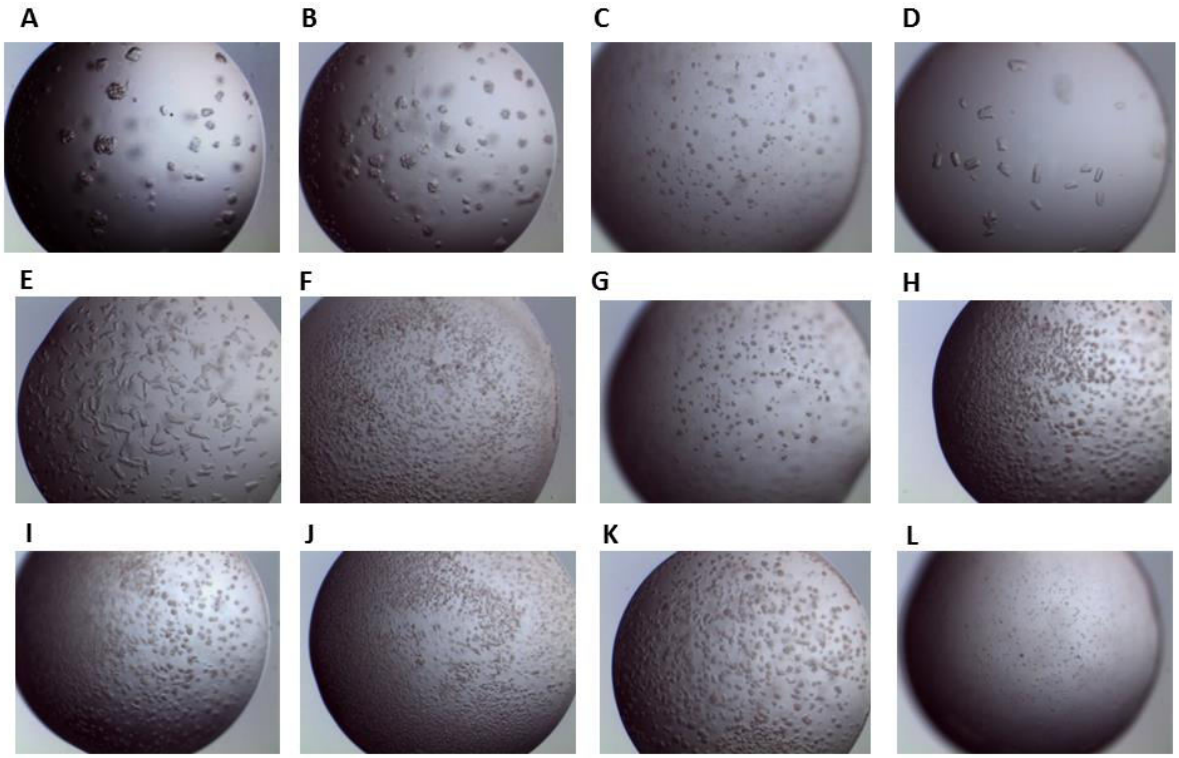


Figure 5.2: Results of additive screen on EcoP1I-SF-DNA. Pictures of crystals in a sitting drop 48 well crystallization plate with A] Potassium chloride, B] Lithium chloride, C] Sucrose, D] NDSB 195, E] NDSB 201, F] NDSB 256, G] Glucose monohydrate, H] Sodium bromide, I] EDTA, J] PEG 3350, K] Pentanerythrol ethoxylate, L] Hexamine cobalt (III) chloride as additive to PC . The drop size is 200 nL. The images captured by Leica S8AP0 with 4X zoom.

Further, to get single, bigger crystals, we tested combinations of different additives with parent condition 2 (Table 5.2). Crystals of various morphologies appeared after 3-4 days. We obtained single crystals (> 100 μm) with M1, M3, M4, M5, M6, M10 and M11 (Figure 5.3, A, C-H). Clusters of crystals could be seen with M2 (Figure 5.3, B). With M7, M8 and M9, tiny plate-like crystals (~50 μm) could be seen (data not shown). To examine the quality of these crystals, we collected a couple of frames at the in-house X-ray diffractometer. Crystals from M2, M3, M4, M5, M6 and M10 diffracted poorly with varying resolutions, all lesser than 7 Å. A crystal from M10 diffracted to better than 6.5 Å in-house. Data collection and processing statistics for this crystal are given in Table 5.4.

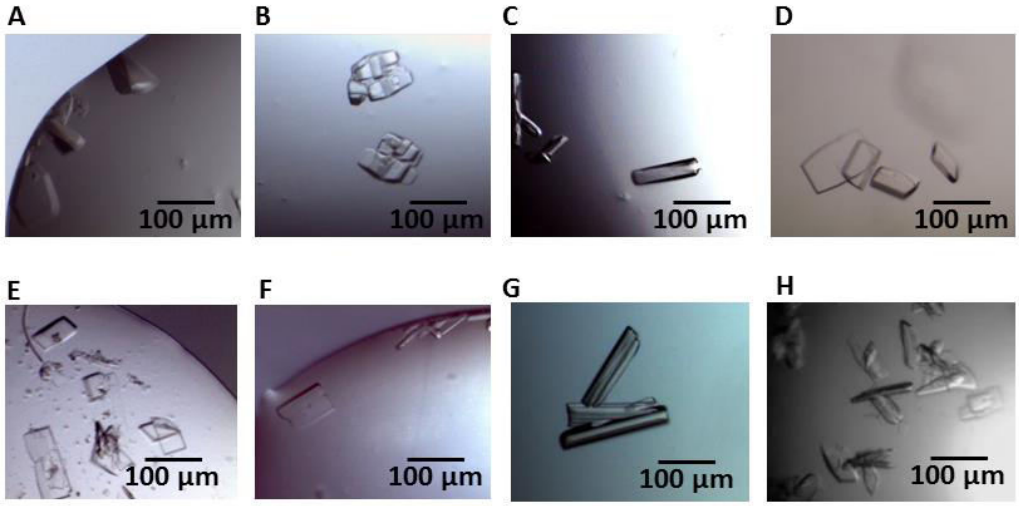


Figure 5.3: Optimization of co-crystallization of EcoP1I-DNA. Pictures of crystals with a variety of crystallization cocktails in a sitting drop 48 well crystallization plate: A] M1; B] M2; C] M3; D] M4; E] M5; F] M6; G] M10; H] M11. The drop size is 2 µL. The images are captured by Leica S8AP0 with 8X zoom.

Table 5.4: Data collection and processing statistics I

Space group	I2
a (Å)	124.8
b (Å)	100.6
c (Å)	253.2
b (°)	93.8
Resolution (Å)*	30-6.5 (7.29-6.5)
Total no of unique reflections*	13970 (3804)
Completeness (%)*	85.2 (85.5)
Mean I/ σ (I)*	8.5 (1.4)
R_{merge} (%)*	9.0 (52.5)
No of molecules in the asymmetric unit	1
Matthew's Coefficient	2.77
Solvent content (%)	55.63

*Values in parentheses denote the last resolution shell.

3.2 Variation in lengths of DNA substrate mimic

All the above optimization trials were done for a 36 bp DNA substrate mimic. This DNA substrate has 5 bp upstream and 26 bp downstream of the target site of EcoP1I (Figure 5.4, A). Earlier DNA binding studies with EcoP1I demonstrated that variation in length upstream of the target site did not affect DNA binding affinity of EcoP1I (Figure 2.8, Chapter 2). To further explore the success of co-crystallization, along with 5/26_P1, we designed two DNA substrates, 6/26_P1 and 7/26_P1 with an upstream length of 6 and 7 bp respectively (Figure 5.4, B and C). We obtained crystals with both of these DNA

substrate mimics in M10. After screening many crystals, few crystals with 6/26_P1 diffracted to better than 5 Å, while those with 7/26_P1 diffracted to better than 6 Å in-house. One of the crystals with 6/26_P1 diffracted to better than 4.5 Å at the synchrotron. Though anisotropic, we collected a complete dataset for this crystal (Table 5.5).

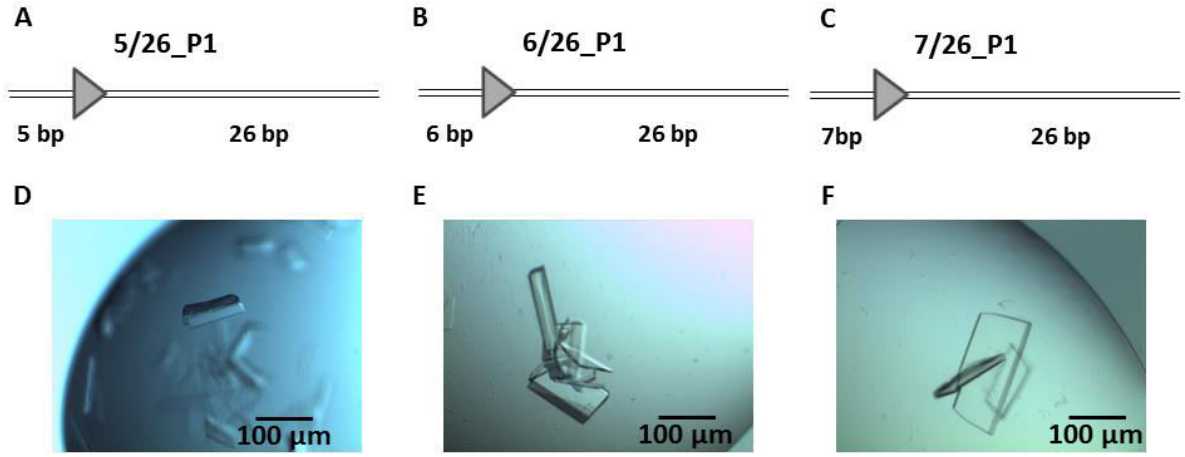


Figure 5.4: Co-crystallization trials with different DNA substrate mimics. Schematic of DNA substrate mimics: A] 5/26_P1, B] 6/26_P1, C] 7/26_P1 in a sitting drop 48 well crystallization plate. D-E] Pictures of crystals with 5/26_P1, 6/26_P1 and 7/26_P1 respectively in a sitting drop 48 well crystallization plate .The drop size is 2 µL. The images are captured by Leica S8AP0 with 8X zoom.

Table 5.5: Data collection and processing statistics II

Space group	I2
a (Å)	125.1
b (Å)	101.8
c (Å)	255.9
b (°)	94.3
Resolution (Å)*	50-4.4 (4.8-4.4)
Total no of unique reflections*	20326 (4829)
Completeness (%)*	98.8 (98.8)
Mean I/ σ (I)*	5.4 (1.4)
R_{merge} (%)*	6.7 (77.3)
No of molecules in the asymmetric unit	1
Matthew's Coefficient	2.77
Solvent content (%)	55.63

*Values in parentheses denote the last resolution shell.

3.3 Experimental phase determination

Before initiating efforts towards experimental phase determination, we counted the number of methionines in EcoP1I in each functional domain. Table 4.8 shows the distribution of methionines in EcoP1I and their locations in partial structure of EcoP15I (PDB ID: 4ZCF). Mod has 4 methionine residues while Res has 14 methionine residues. Interestingly, there are 5 methionine residues in the flexible linker between ATPase and endonuclease domains. To determine phases experimentally, we purified selenomethionine (Se-Met) derivative of EcoP1I$^{N\text{-}His}$. The Se-Met derivative of EcoP1I$^{N\text{-}His}$ was crystallized with a 5/26_P1 in presence of SF. One of such crystals diffracted to less than 5 Å at IO4, DLS. Since the data was highly anisotropic, a complete dataset could not be collected.

Table 5.6: Distribution of methionine residues in EcoP1I

Domain	*Structural boundaries of EcoP15I based on 4ZCF*	*No of methionines in EcoP1I*	*Primary seq no in EcoP1I*
Mod: 4 Methionines			
NTD	**K2-T76**	**1**	**1**
Catalytic Domain	**N92-S260**	**3**	**109,119,234**
TRD	**N261- L394**		**-**
AdoMet Binding	**D395-G506**		**-**
CTD	**F507- N644**		**-**
Res: 14 Methionines			
RecA1	**T6-L267**	**5**	**1,22,84,163,190**
RecA2 including Pin	**T268- I594**	**4**	**374,375,544,607**
RecA2 before Pin	**T268-D287**	**0**	**-**
after Pin	**P363-I594**	**4**	**374,375,544,607**
Pin	**N293-N362**	**0**	**-**
Linker between motor and nuclease	**D620-L810**	**5**	**660,699,724,726,743**
Nuclease	**Till P970**	**0**	**-**

3.4 Partial structure solution of EcoP1I

The first X-ray diffraction data from an EcoP1I-DNA crystal was collected in-house. This data, which was indexed and processed using XDS (19) and scaled and merged using AIMLESS (20), diffracted to 6.5 Å (Table 5.4). A similar crystal diffracted to 4.4 Å at the I04 beamline of DLS (Figure 5.5A). This data was indexed and processed using MOSFLM (18), and scaled and merged using AIMLESS (Table 5.5). The better resolved data was used for MR using the program PHASER (22). Using the coordinates of EcoP15I a partial structure solution was obtained. The solution was clear and had a high LLG value of 210 and TFZ value of 12.2. Figure 5.5B shows the packing of the subunits in the crystal

lattice. The model thus obtained was refined using PHENIX (23). Due to the low resolution of the data, only a single round of rigid body refinement was carried out. B-factors of all the atoms were set to 75 $Å^2$. R and R_{free} after refinement were 48.3 % and 47.6 %, respectively. The electron density map calculated using the coordinates of the refined partial model was viewed using COOT (24). The map revealed discontinuous but discernible electron density of the bound DNA. However, the density for the nuclease domain was not clear.

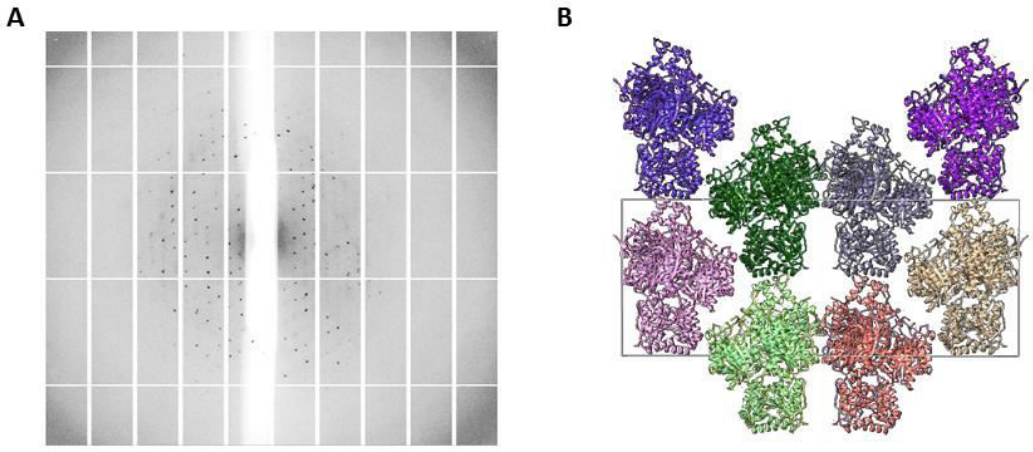


Figure 5.5: Crystallographic studies of EcoP1I-DNA complex. A] X-ray diffraction pattern of EcoP1I-DNA crystal showing intensities of spots. B] Packing of the subunits in the crystal lattice of cell I2. The unit cell is depicted using grey lines.

4. Discussion

To understand the mechanism of working of Type III RM enzymes, we initiated crystallographic studies with EcoP1I. The enzyme forms 260 kDa assembly of 2 Mods and 1 Res. We aimed at co-crystallizing EcoP1I-DNA complex. Crystallizing such a huge protein-DNA complex was a not a trivial task. We used EcoP1I$^{N\text{-}His}$ and EcoP1I (E916A)$^{C\text{-}His}$ for crystallization experiments.

Design of DNA substrate mimics: As an initial step towards co-crystallization, we characterized the DNA binding properties of EcoP1I with a variety of DNA substrates and buffer components (See Chapter 2). By systematically studying the effects of DNA length on DNA binding and endonucleolytic activity of EcoP1I, we designed appropriate oligomers for co-crystallization experiments. Heterologous cooperation assays suggested that minimum 17 bp was necessary for engaging the ATPase domain on DNA, while cleavage assays with 5/25_P1 and 5/24_P1 suggested that an oligomer with 24 bp downstream of target site was refractory to dsDNA break. These results gave important insights about effect on activity of ATPase and endonuclease domains with various DNAs. Based on all these experiments, we designed three DNA substrate mimics for co-crystallization studies, namely 5/26_P1, 6/26_P1 and 7/26_P1.

The EcoP15I-DNA structure published recently does not have information about the endonuclease domain and the linker between endonuclease and ATPase (3). As can be seen from the structure, the RecA2 subdomain is wheeled away from DNA. We suspect that this is because of short length of DNA substrate mimic used for setting up crystallization. With careful and systematic approach of designing the oligomers suitable for co-crystallization with EcoP1I, we are confident that in our crystallization experiments the enzyme is fully engaged with DNA.

Optimization of crystallization conditions: We screened a variety of conditions for crystallization. Most promising crystals were obtained in a condition composed of PEG 6000, MPD and Na-HEPES (pH 7.5). Another important factor during co-crystallization of EcoP1I-DNA was to choose the most optimum storage buffer for EcoP1I. The final step of purification of EcoP1I was SEC, which was carried out using Buffer D, a potassium based buffer. An earlier experiment suggested that DNA binding affinity of

EcoP1I was unaltered in presence of either potassium or sodium ions (Section 3.3.1, Chapter 2). We compared the quality of crystals by replacing potassium ions with sodium ions in final storage buffer. Presence of sodium ions in storage buffer resulted in chunkier crystals. Further, we examined effect of various additives on crystal quality. We observed crystals with 21 various additives. Addition of most of the additives resulted in clusters of crystals and micro-crsytals; however, we obtained tiny single crystals with hexamine cobalt (III) chloride (Werner's Reagent). We further combined various additives to improve the size and quality of crystals. A combination of Werner's reagent and NDSB 201 gave single chunky crystals measuring up to 300 μm. To assess the quality of crystals, we screened them at in-house X-ray diffraction facility. After screening a large number of crystals, one of them diffracted to 6.5 Å. A similar crystal diffracted to 4.4 Å at the synchrotron. We further went ahead with molecular replacement by using EcoP15I-DNA structure (PDB ID: 4ZCF) as the initial search model (3).

The discovery of Type III RM enzymes dates back to 1965. Complete structural information about any of the prototype has been evasive due to large size of the complex and flexibility associated with various domains within. We took up the challenge to co-crystallize EcoP1I with a suitable DNA substrate mimic. We succeeded in crystallizing the EcoP1I-DNA complex with three different DNA substrate mimics. Despite getting good quality crystals, many were damaged during shipment to synchrotron. One of the crystals of EcoP1I-6/26_P1 diffracted to 4.4 Å. However, the diffraction data that we have is anisotropic. The quality of the map is not good enough to be able to solve the complete structure. To solve the complete structure of EcoP1I-DNA complex, we need phase information and crystals diffracting at even better resolution. Soaking experiments carried out with tantalum bromide and sodium tungstate decreased the diffraction quality of original crystals, and also resulted in significant changes in cell parameters (data not shown). Crystals of selenomethionine derivative of EcoP1I have been obtained. Work is in progress to obtain higher resolution data with anomalous signal, which can be used to phase using the method of single-wavelength anomalous diffraction with partial molecular replacement model.

5. References

1. Wyszomirski,K.H., Curth,U., Mackeldanz,P., Schutkowski,M., Kru,D.H. and Mo,E. (2011) Type III restriction endonuclease EcoP15I is a heterotrimeric complex containing one Res subunit with several DNA-binding regions and ATPase activity. *Nucleic Acids Res.*, 10.1093/nar/gkr1239.

2. Butterer,A., Pernstich,C., Smith,R.M., Sobott,F., Szczelkun,M.D. and Tóth,J. (2014) Type III restriction endonucleases are heterotrimeric: comprising one helicase-nuclease subunit and a dimeric methyltransferase that binds only one specific DNA. *Nucleic Acids Res.*, **42**, 5139–50.

3. Gupta,Y.K., Chan,S.-H., Xu,S. and Aggarwal,A.K. (2015) Structural basis of asymmetric DNA methylation and ATP-triggered long-range diffusion by EcoP15I. *Nat. Commun.*, **6**, 1–10.

4. Timinskas,A., Butkus,V. and Janulaitis,A. (1995) Sequence motifs characteristic for DNA [cytosine-N4] and DNA [adenine-N6] methyltransferases . Classification of all DNA methyltransferases. *Gene*, **157**, 3–11.

5. Malone,T., Blumenthal,R.M., Cheng,X. and Structural,W.M.K. (1995) Structure-guided analysis reveals nine sequence motifs conserved among DNA amino-methyl-transferases , and suggests a catalytic mechanism for these enzymes. *J. Mol. Biol.*, **253**, 618–632.

6. Gorbalenya,A.E. and Koonin,E. V (1991) Endonuclease (R) subunits of type-I and type -III restriction-modication enzymes contain a helicase-like domain. *FEBS Lett.*, **291**, 277–281.

7. Aravind,L., Makarova,K.S. and Koonin,E. V (2000) Holliday junction resolvases and related nucleases : identification of new families , phyletic distribution and evolutionary trajectories. *Nucleic Acids Res.*, **28**, 3417–3432.

8. Wagenführ,K., Pieper,S., Mackeldanz,P., Linscheid,M., Krüger,D.H. and Reuter,M. (2007) Structural domains in the type III restriction endonuclease EcoP15I: characterization by limited proteolysis, mass spectrometry and insertional mutagenesis. *J. Mol. Biol.*, **366**, 93–102.

9. Fairman-williams,M.E., Guenther,U. and Jankowsky,E. (2010) SF1 and SF2 helicases : family matters. *Curr. Opin. Struct. Biol.*, **20**, 313–324.

10. Brockes,J.P. (1973) The deoxyribonucleic acid-modification enzyme of bacteriophage P1. *Biochem. J.*, **133**, 629–633.

11. Hadi,S.M., Baechi,B., Shepherd,J.C.W., Yuan,R., Ineichen,K. and Bickle,T.A. (1979) DNA Recognition and cleavage by the EcoPl5 restriction endonuclease. *J. Mol. Biol.*, **134**, 655–666.

12. Toth,J., Bollins,J. and Szczelkun,M.D. (2015) Re-evaluating the kinetics of ATP hydrolysis during initiation of DNA sliding by Type III restriction enzymes. *Nucleic Acids Res.*, **43**, 10870–10881.

13. Meisel,A., Bicklel,T.A., Kruger,D.H. and Schroeder,C. (1992) Type III restriction enzymes need two inversely oriented recognition sites for DNA cleavage. *Nature*, **355**, 467–469.

14. Schwarz,F.W., Toth,J., van Aelst,K., Cui,G., Clausing,S., Szczelkun,M.D. and Seidel,R. (2013) The helicase-like domains of type III restriction enzymes trigger long-range diffusion along DNA. *Science (80-.).*, **340**, 353–356.

15. Smith,R.M., Josephsen,J. and Szczelkun,M.D. (2009) The single polypeptide restriction – modification enzyme LlaGI is a self-contained molecular motor that translocates DNA loops. *Nucleic Acids Res.*, **37**, 7219–7230.

16. Murray,N.E. (2000) Type I restriction systems: Sophisticated molecular machines (a Legacy of Bertani and Weigle). *Microbiol. Mol. Biol. Rev.*, **64**, 412–434.

17. Aelst,K. Van, Tóth,J., Ramanathan,S.P., Schwarz,F.W., Seidel,R. and Szczelkun,M.D. (2010) Type III restriction enzymes cleave DNA by long-range interaction between sites in both head-to-head and tail-to-tail inverted repeat. *PNAS*, **107**, 9123–9128.

18. Leslie,A. and Powell,H. (2007) Evolving methods for macromolecular crystallography. In *Springer*.pp. 41–51.

19. Kabsch,W. (2010) XDS. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **D66**, 125–132.

20. Evans,P.R. and Murshudov,G.N. (2013) How good are my data and what is the

resolution? *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **D69**, 1204–1214.

21. Evans,P. (2006) Scaling and assessment of data quality. *Acta Cryst*, **62**, 72–82.

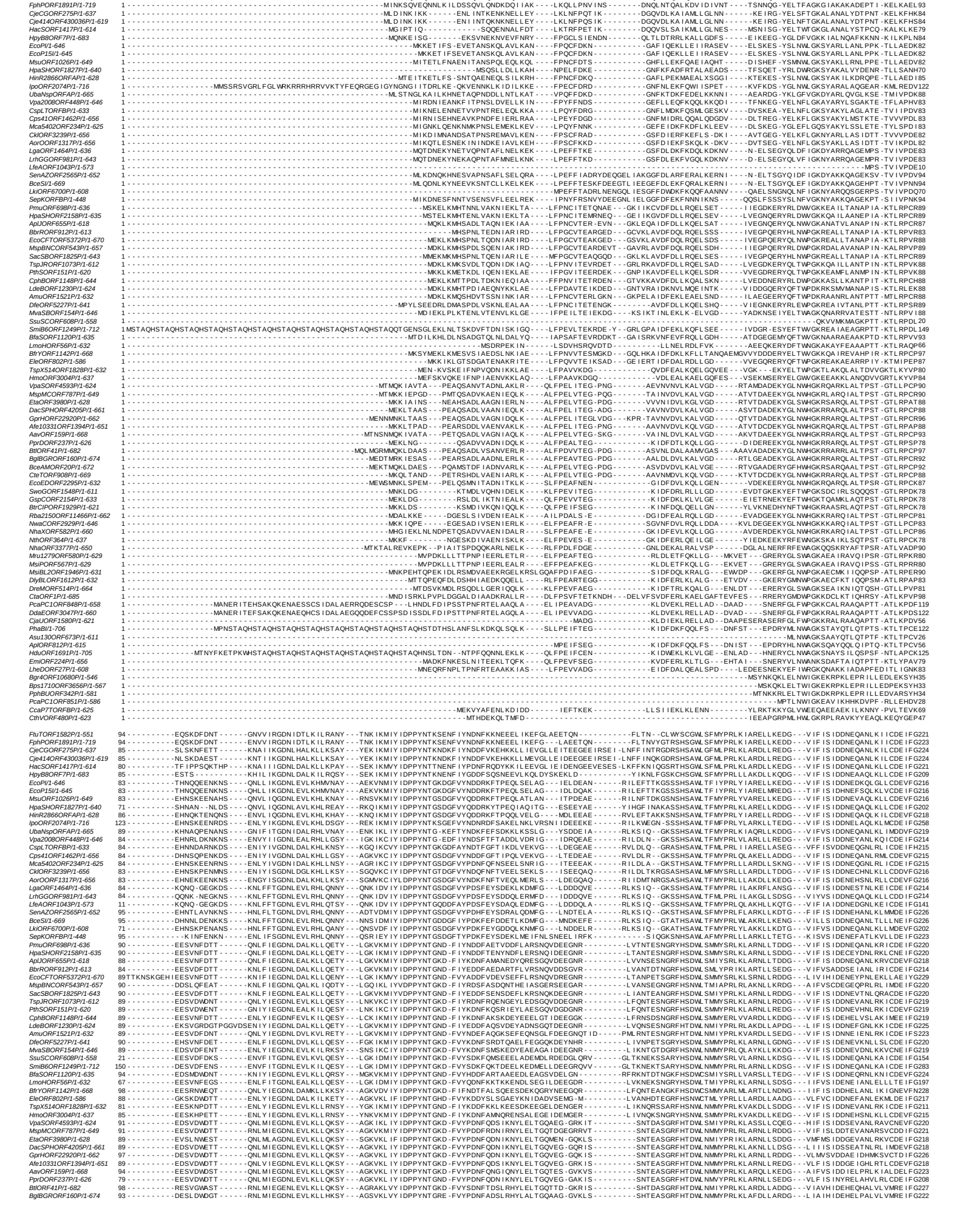
22. Read,R.J. and McCoy,A.J. (2011) Using SAD data in Phaser. *Acta Crystallogr. D. Biol. Crystallogr.*, **67**, 338–344.

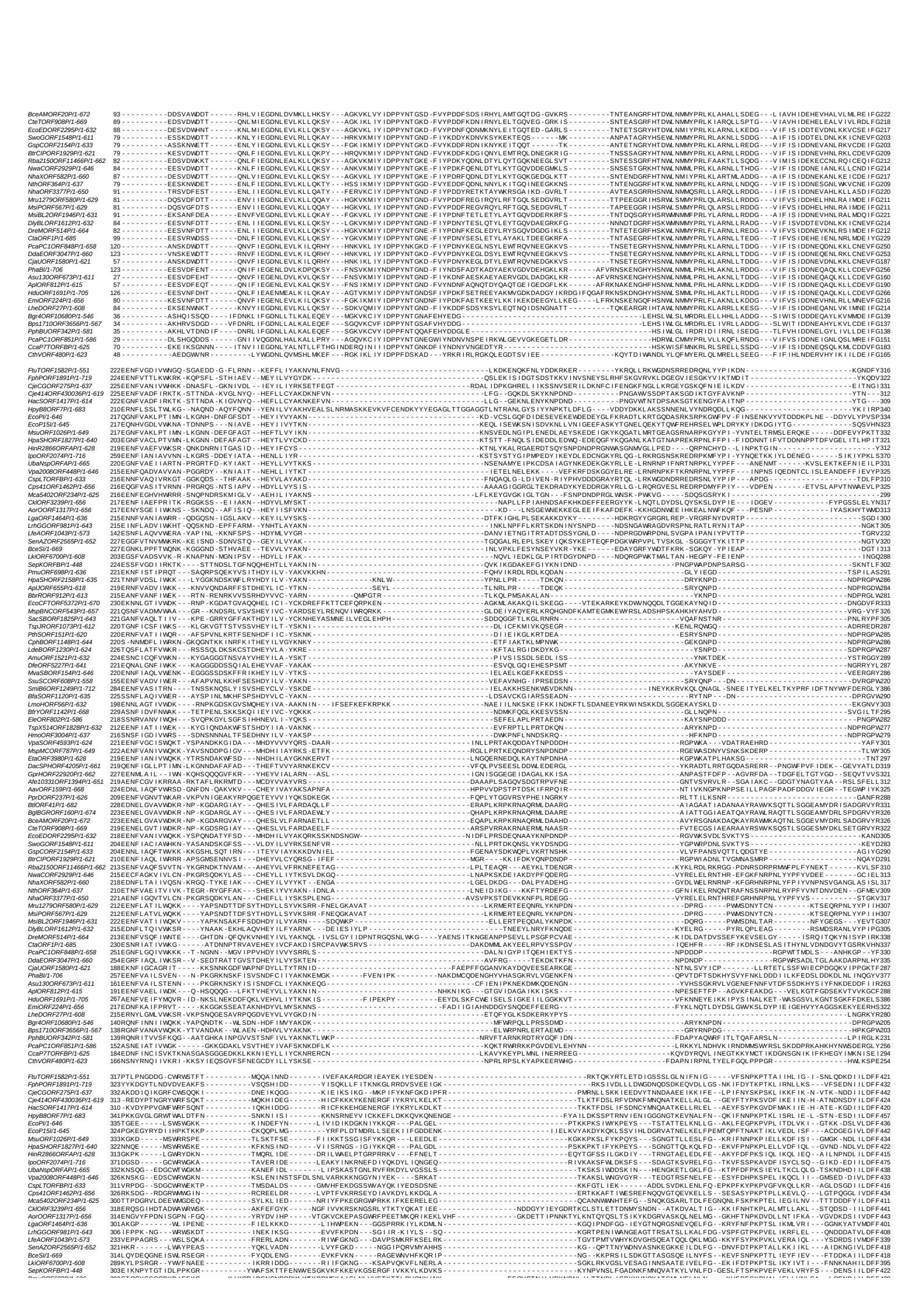
23. Adams,P.D., Afonine,P. V, Bunkóczi,G., Chen,V.B., Davis,I.W., Echols,N., Headd,J.J., Hung,L.-W., Kapral,G.J., Grosse-Kunstleve,R.W., *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **66**, 213–221.

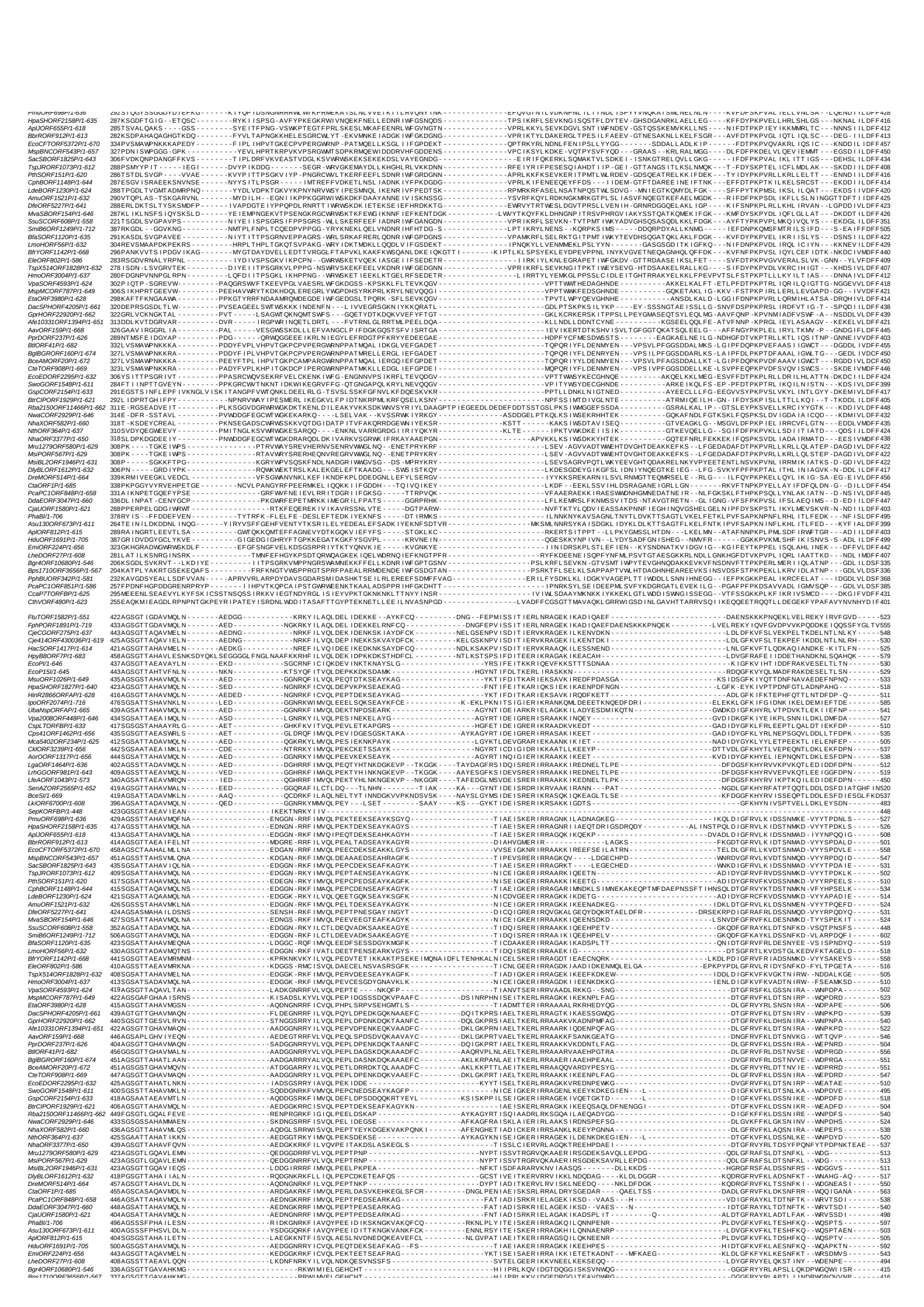
24. Emsley,P., Lohkamp,B., Scott,W.G. and Cowtan,K. (2010) Features and development of Coot. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **66**, 486–501.

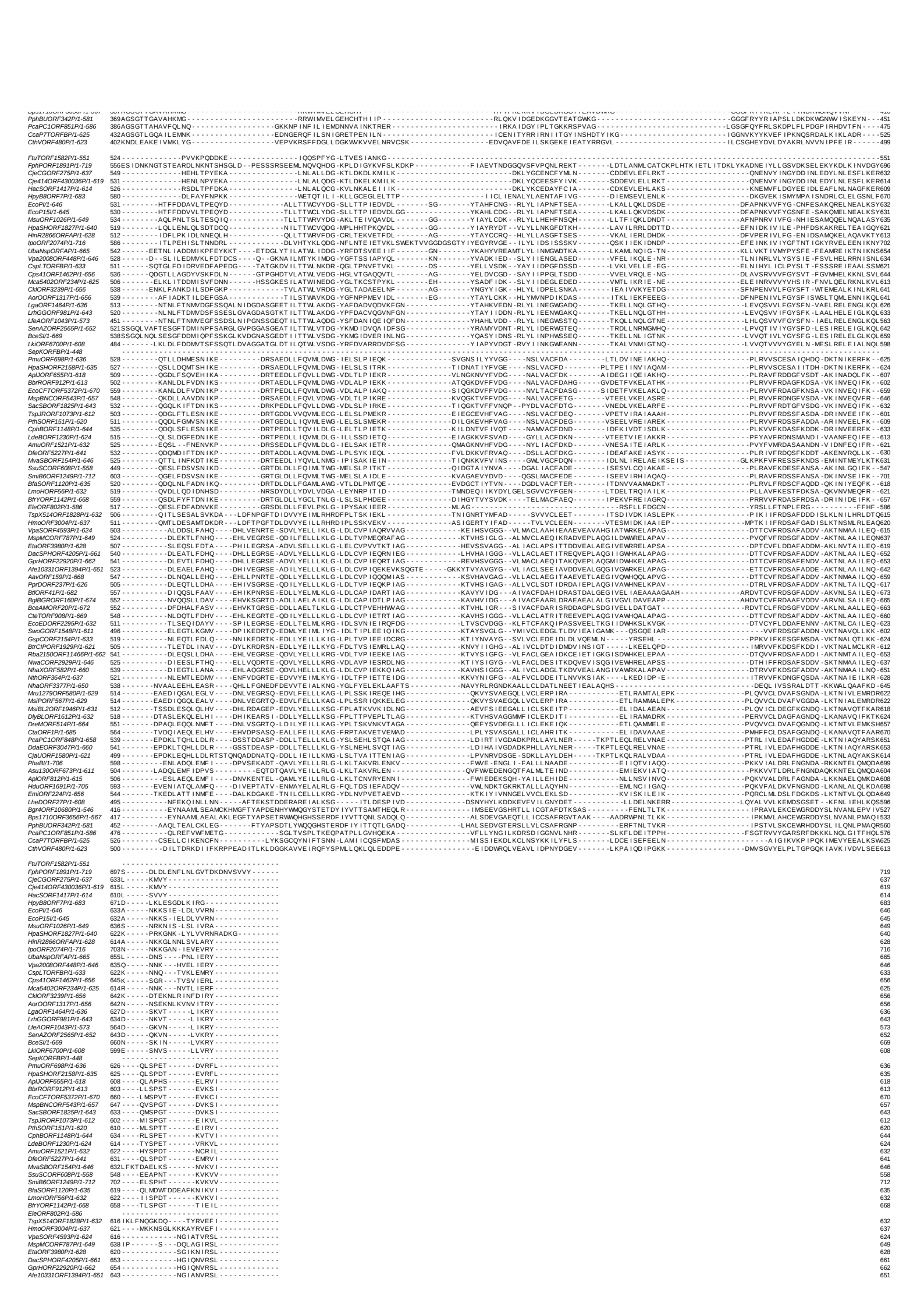
APPENDIX I

Type III MTase









```
AavORF159P/1-668          660 ----------HGIQNVRSL---------- 668
PprDORF237P/1-626         618 ----------QGLDDVRSL---------- 626
BtlORF41P/1-682           674 ----------HGVKRIRSL---------- 682
BglBGRORF160P/1-674       666 ----------HGIRSVRSL---------- 674
BceAMORF20P/1-672         664 ----------HGVKRVRSL---------- 672
CteTORF908P/1-669         661 ----------HGIANVRSL---------- 669
EcoEDORF2295P/1-632       624 ----------HGLTNLHSL---------- 632
SwoGORF1548P/1-611        603 ----------FGIEDFRTI---------- 611
GspCORF2154P/1-633        625 ----------HGITDVRSV---------- 633
BtrCIPORF1929P/1-621      613 ----------YGIADVKSL---------- 621
Rba2150ORF11466P/1-662    654 ----------NGIAHVRSL---------- 662
NwaCORF2929P/1-646        638 ----------NGINHVRSL---------- 646
NhaXORF582P/1-660         652 ----------NGILDVRSL---------- 660
NthORF364P/1-637          629 ----------HGIEEIMSI---------- 637
NhaORF3377P/1-650         646 ----------NLHTA---------- 650
Mru1279ORF580P/1-629      623 G----------IRFRTA---------- 629
MsiPORF567P/1-629         623 G----------IRFRTV---------- 629
MsiBL2ORF1946P/1-631      619 AQ------AEE--SEIVFKTV---------- 631
DlyBLORF1612P/1-632       625 EI------K--------KFETV---------- 632
DreMORF514P/1-664         658 KI------I--------FHTV---------- 664
CtaORF1P/1-685            671 NQ------GREKASQIVFRTV---------- 685
PcaPC1ORF848P/1-658       652 G----------IELRTA---------- 658
DdaEORF3047P/1-660        654 G----------IELRTA---------- 660
CjaUORF1580P/1-621        615 G----------IELRTH---------- 621
PhaBI/1-706               700 G----------ITFFVI---------- 706
Asu130ORF673P/1-611       605 D----------IAFFVI---------- 611
AplORF812P/1-615          609 N----------IAFFVI---------- 615
HduORF1691P/1-705         699 G----------IAFETF---------- 705
EmiORF224P/1-656          650 E----------IDLTVV---------- 656
LheDORF27P/1-608          597 AN---------LMNHFNLEWI---------- 608
Bgr4ORF10680P/1-546       528 VK----------ENNKNKDIVQAELD--LFE--- 546
Bps1710ORF3656P/1-567     534 EKAQAVVTGAGIKKTKSKSAARNERQGGLFGENE 567
PphBUORF342P/1-581        561 EEETLKKKKSAHPSLFDNQDA---------- 581
PcaPC1ORF851P/1-586       577 P----------YELAVKTWL---------- 586
CcaP7TORFBP/1-625             ---------- ----------
CthVORF480P/1-623         614 Y----------FEVLEVGAR---------- 623
```

Table A4 : Evolutionary conservation scores (ConSurf grades) of MTase*

*The table details the residue variety in % for each position in the query sequence.
Each column shows the % for that amino-acid, found in position ('pos') in the MSA.

pos	ConSurf Grade	pos	ConSurf Grade	pos	ConSurf Grade	pos	ConSurf Grade	pos	ConSurf Grade	pos	ConSurf Grade
1	4	60	9	119	8	178	7	237	2	296	4
2	2	61	8	120	9	179	8	238	6	297	3
3	5	62	9	121	9	180	7	239	1	298	5
4	7	63	9	122	9	181	9	240	1	299	4
5	6	64	7	123	9	182	9	241	8	300	5
6	3	65	7	124	9	183	6	242	6	301	1
7	6	66	9	125	9	184	7	243	6	302	1
8	6	67	7	126	9	185	9	244	5	303	3*
9	7	68	6	127	9	186	7	245	5	304	4
10	6	69	5	128	9	187	8	246	2	305	5
11	8	70	8	129	9	188	9	247	7	306	7
12	7	71	9	130	7	189	9	248	3	307	4
13	5	72	7	131	9	190	3	249	8	308	1
14	6	73	8	132	7	191	4	250	4	309	7
15	5	74	8	133	9	192	8	251	1	310	7
16	3	75	6	134	8	193	9	252	1	311	1
17	9	76	8	135	9	194	6	253	1	312	5
18	7	77	9	136	5	195	8	254	1	313	9
19	4	78	9	137	9	196	7	255	1	314	1
20	6	79	7	138	7	197	8	256	1	315	4
21	8	80	9	139	8	198	9	257	3	316	1
22	6	81	4	140	5	199	9	258	1	317	1
23	4	82	5	141	6	200	9	259	2	318	1
24	7	83	5	142	6	201	8	260	2	319	2
25	9	84	8	143	3	202	5	261	1	320	2
26	9	85	9	144	3	203	9	262	7	321	2
27	6	86	7	145	6	204	2	263	4	322	1
28	6	87	6	146	7	205	2	264	1	323	1
29	6	88	8	147	6	206	8	265	1	324	1
30	8	89	7	148	3	207	8	266	1	325	1
31	9	90	9	149	4	208	5	267	1	326	1
32	4	91	7	150	8	209	4	268	6	327	1
33	8	92	5	151	6	210	4	269	1	328	2
34	7	93	9	152	8	211	4	270	6	329	1
35	7	94	6	153	1	212	6	271	1	330	1
36	8	95	9	154	7	213	9	272	7	331	7
37	7	96	7	155	5	214	7	273	1	332	1
38	7	97	9	156	4	215	9	274	1	333	6
39	6	98	9	157	8	216	9	275	1	334	1
40	9	99	9	158	5	217	6	276	2	335	2
41	6	100	9	159	7	218	2	277	3	336	8
42	8	101	9	160	6	219	9	278	4*	337	5
43	6	102	9	161	7	220	7	279	2	338	5
44	8	103	9	162	6	221	2	280	1	339	1
45	7	104	8	163	9	222	4	281	2	340	1
46	4	105	9	164	6	223	1	282	4*	341	1
47	2	106	9	165	7	224	4	283	1	342	1
48	6	107	9	166	6	225	5	284	1	343	1
49	6	108	8	167	8	226	6	285	4	344	1
50	6	109	9	168	9	227	3	286	5	345	5
51	5	110	9	169	9	228	4	287	6	346	1
52	5	111	8	170	8	229	3	288	3	347	1
53	6	112	8	171	9	230	1	289	3	348	1
54	9	113	9	172	9	231	1	290	3	349	8
55	8	114	7	173	8	232	7	291	2	350	3
56	9	115	7	174	9	233	5	292	5	351	1
57	7	116	8	175	9	234	3	293	1	352	3
58	9	117	7	176	8	235	2	294	4	353	2
59	7	118	9	177	9	236	1	295	3	354	2

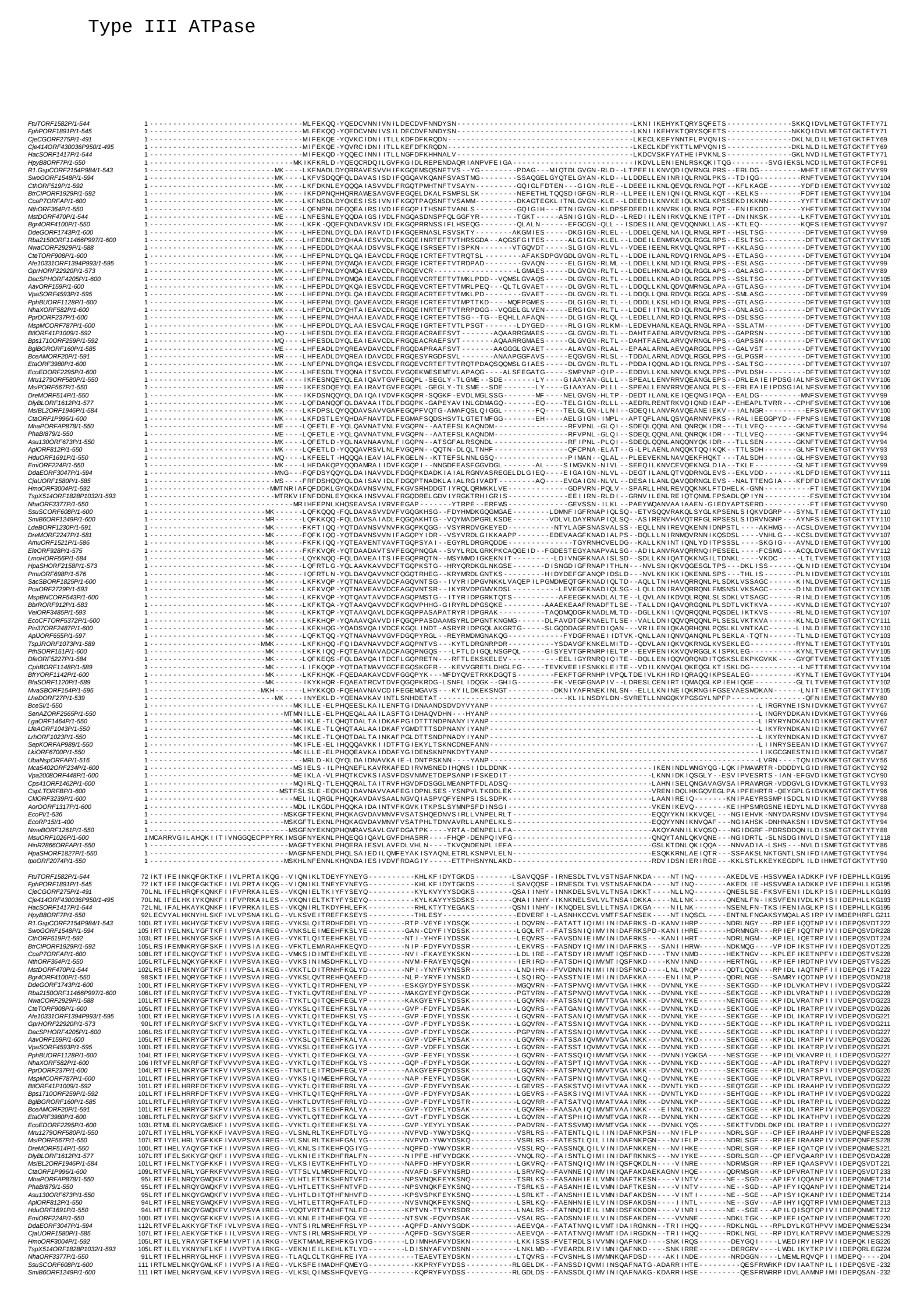
pos	ConSurf Grade	pos	ConSurf Grade	pos	ConSurf Grade	pos	ConSurf Grade	pos	ConSurf Grade
355	1	419	6	483	7	547	7	611	8
356	2	420	1	484	4	548	9	612	9
357	1	421	1	485	2	549	8	613	5
358	1	422	6	486	6	550	7	614	7
359	1	423	4	487	9	551	7	615	8
360	1	424	1	488	7	552	8	616	6
361	1	425	2	489	6	553	7	617	8
362	2	426	1	490	8	554	9	618	6
363	3	427	2	491	7	555	7	619	6
364	3	428	1	492	8	556	6	620	9
365	6	429	2	493	7	557	9	621	7
366	1	430	1	494	9	558	7	622	9
367	6	431	1	495	8	559	6	623	7
368	3	432	1	496	7	560	6	624	7
369	1	433	1	497	7	561	8	625	8
370	1	434	2	498	8	562	5	626	7
371	1	435	4	499	5	563	3	627	8
372	1	436	5	500	3	564	5	628	7
373	1	437	5	501	6	565	1	629	7
374	1	438	7	502	1	566	6	630	6
375	6	439	9	503	2	567	3	631	8
376	1	440	9	504	9	568	5	632	8
377	1	441	9	505	3	569	7	633	3
378	8	442	8	506	9	570	3	634	8
379	1	443	9	507	9	571	8	635	8
380	3	444	9	508	9	572	6	636	5
381	6	445	5	509	7	573	5	637	6
382	2	446	8	510	7	574	8	638	6
383	5	447	8	511	7	575	6	639	7
384	1	448	8	512	7	576	6	640	7
385	1	449	7	513	7	577	8	641	8
386	1	450	9	514	5	578	5	642	7
387	1	451	7	515	8	579	8	643	9
388	1	452	6	516	9	580	8	644	9
389	1	453	4	517	8	581	8		
390	1	454	3	518	5	582	6		
391	1	455	9	519	6	583	2		
392	1	456	4	520	6	584	2		
393	2	457	4	521	6	585	7		
394	2	458	7	522	2	586	5		
395	1	459	1	523	5	587	6		
396	5	460	2	524	4	588	3		
397	3	461	4	525	3	589	4		
398	1	462	8	526	6	590	7		
399	4	463	6	527	1	591	6		
400	3	464	6	528	4	592	4		
401	1	465	8	529	1	593	8		
402	5	466	5	530	3	594	7		
403	3	467	8	531	7	595	4		
404	1	468	9	532	1	596	6		
405	1	469	7	533	5	597	8		
406	6	470	6	534	6	598	9		
407	5	471	9	535	8	599	4		
408	2	472	1	536	6	600	7		
409	1	473	3	537	7	601	7		
410	1	474	2	538	3	602	2		
411	1	475	3	539	5	603	3		
412	1	476	2	540	8	604	4		
413	1	477	7	541	7	605	7		
414	9	478	5	542	4	606	6		
415	4	479	8	543	6	607	8		
416	1	480	1	544	9	608	7		
417	8	481	7	545	8	609	8		
418	9	482	4	546	9	610	7		

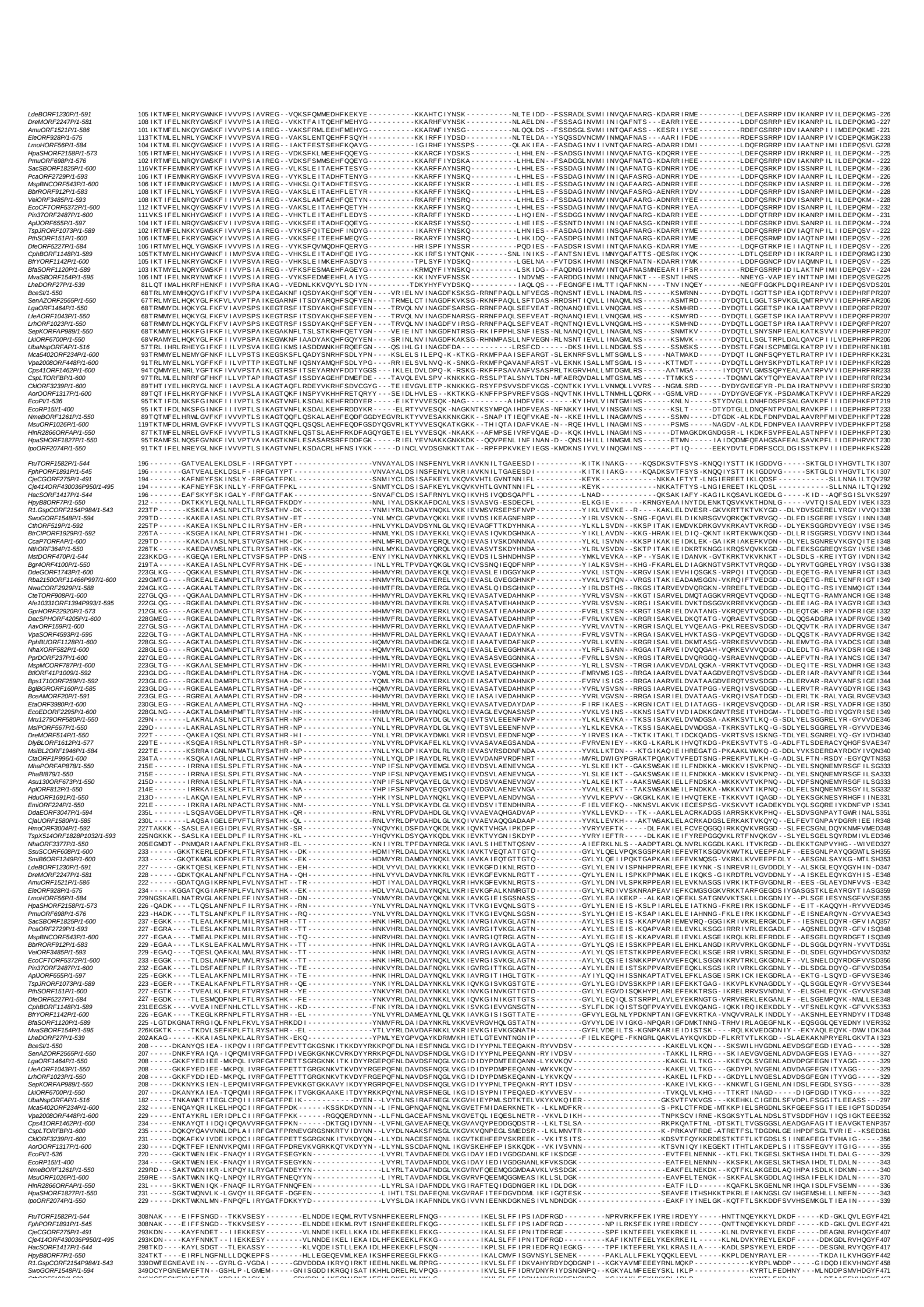
Table A5 : Evolutionary conservation scores (ConSurf grades) of CTD*

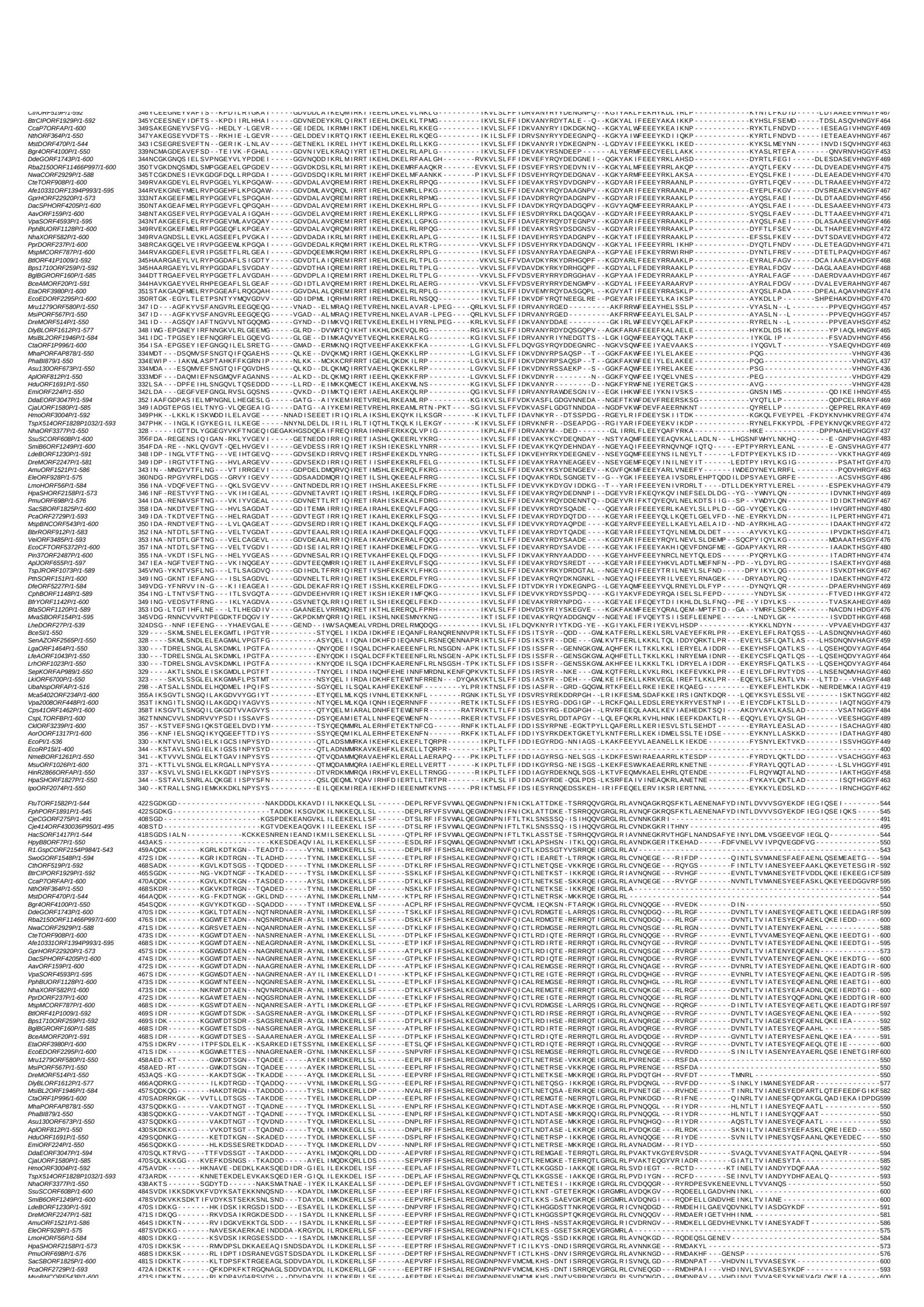
*The table details the residue variety in % for each position in the query sequence.
Each column shows the % for that amino-acid, found in position ('pos') in the MSA.

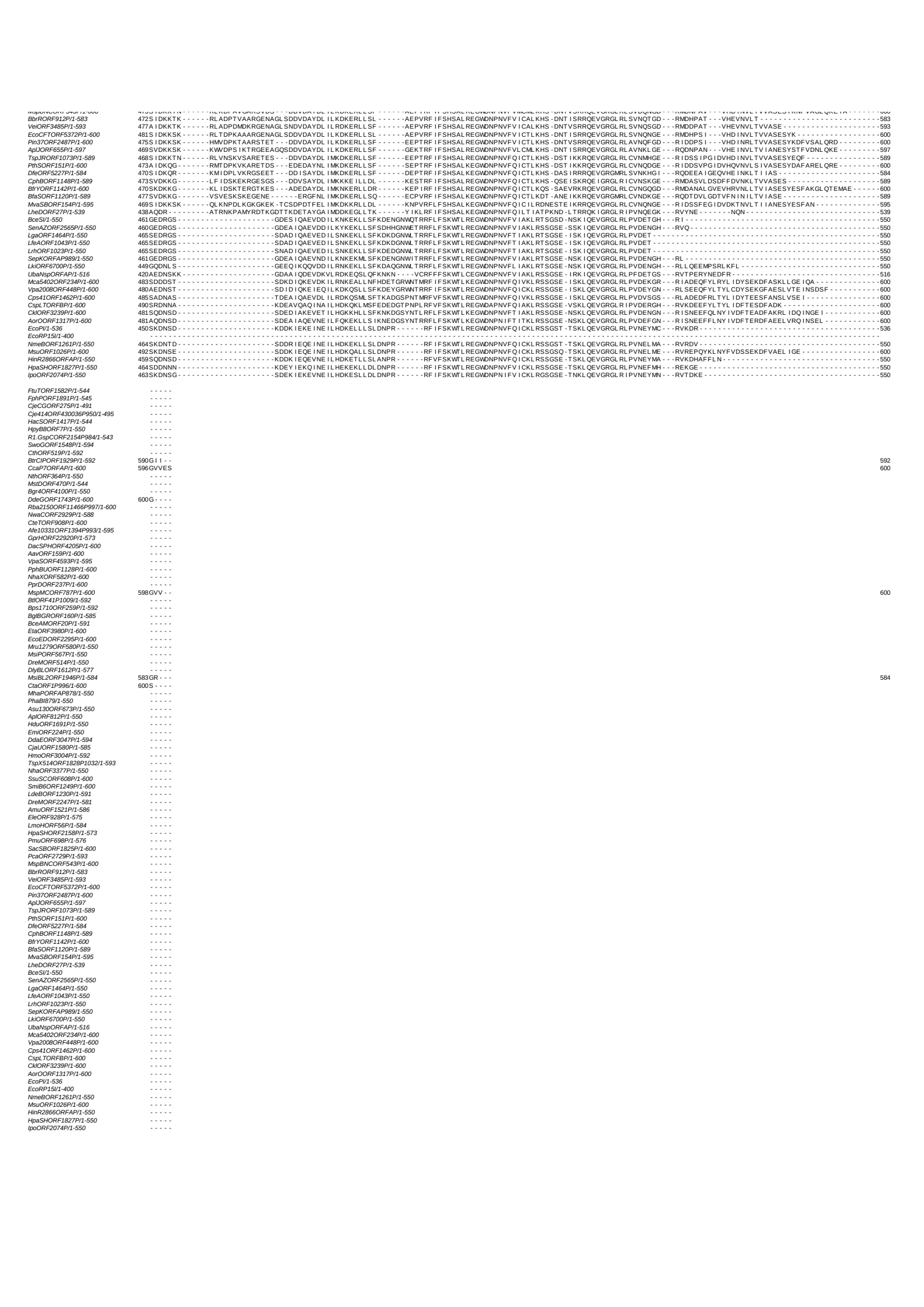
pos	ConSurf Grade	pos	ConSurf Grade	pos	ConSurf Grade	pos	ConSurf Grade	pos	ConSurf Grade	pos	ConSurf Grade
1	9	26	7	51	5	76	8	101	4	126	7
2	8	27	1	52	9	77	4	102	7	127	7
3	9	28	3	53	7	78	1	103	6	128	1
4	5	29	4	54	4	79	1	104	8	129	8
5	5	30	7	55	4	80	5	105	6	130	8
6	6	31	5	56	6	81	3	106	7	131	4
7	6	32	7	57	3	82	4	107	9	132	5
8	6	33	1	58	1	83	1	108	4	133	5
9	3	34	3	59	2	84	3	109	7	134	6
10	8	35	8	60	1	85	6	110	8	135	6
11	9	36	6	61	4	86	5	111	4	136	8
12	7	37	2	62	1	87	1	112	9	137	7
13	4	38	5	63	3	88	7	113	5	138	9
14	4	39	9	64	6	89	6	114	5	139	8
15	5	40	7	65	1	90	3	115	9		
16	4	41	8	66	7	91	4	116	7		
17	1	42	7	67	5	92	7	117	9		
18	2	43	9	68	3	93	8	118	7		
19	2	44	7	69	7	94	3	119	5		
20	1	45	7	70	5	95	6	120	8		
21	4	46	7	71	4	96	5	121	6		
22	1	47	8	72	8	97	1	122	8		
23	2	48	7	73	3	98	1	123	6		
24	1	49	9	74	8	99	2	124	6		
25	1	50	6	75	7	100	7	125	4		

Type III ATPase

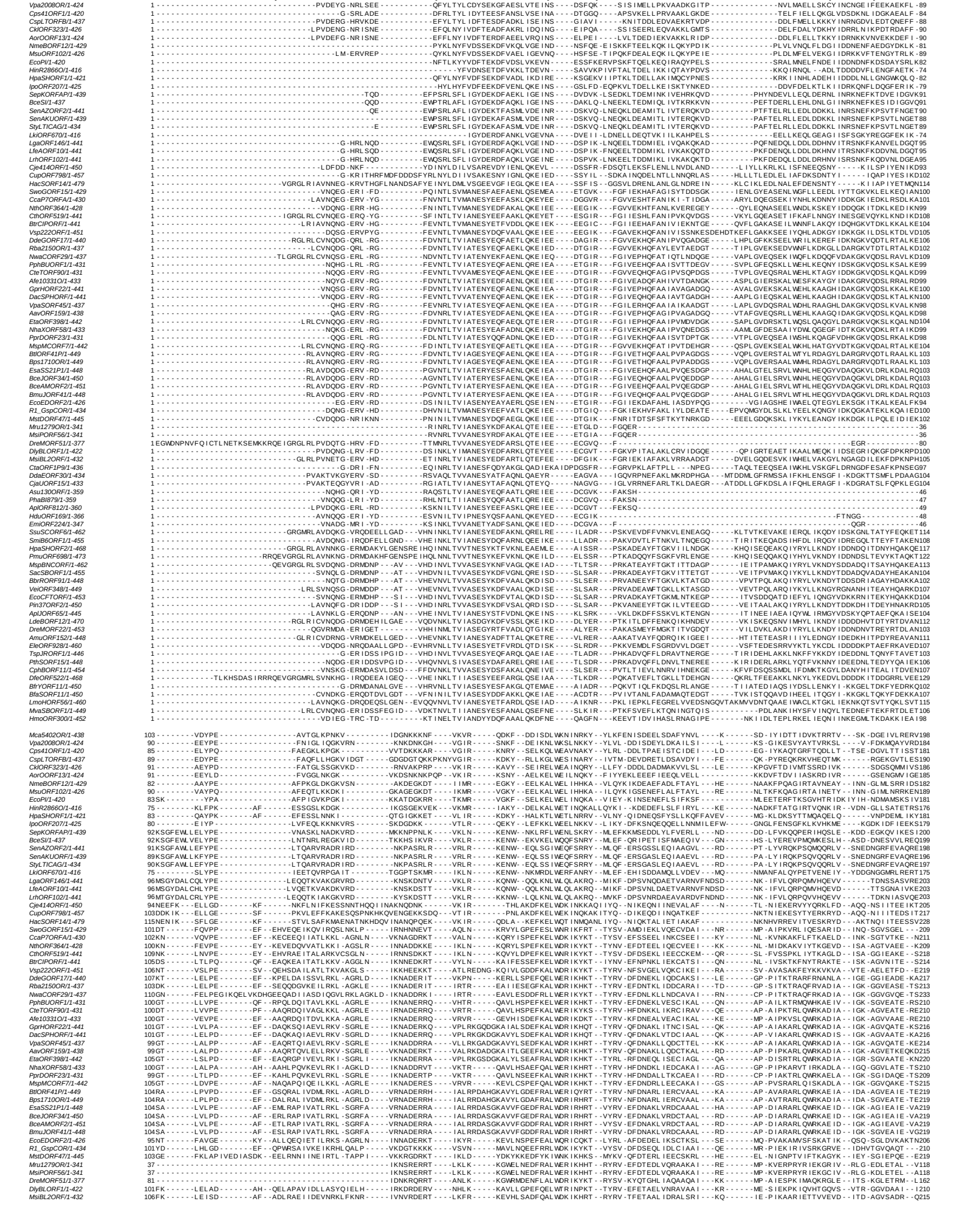


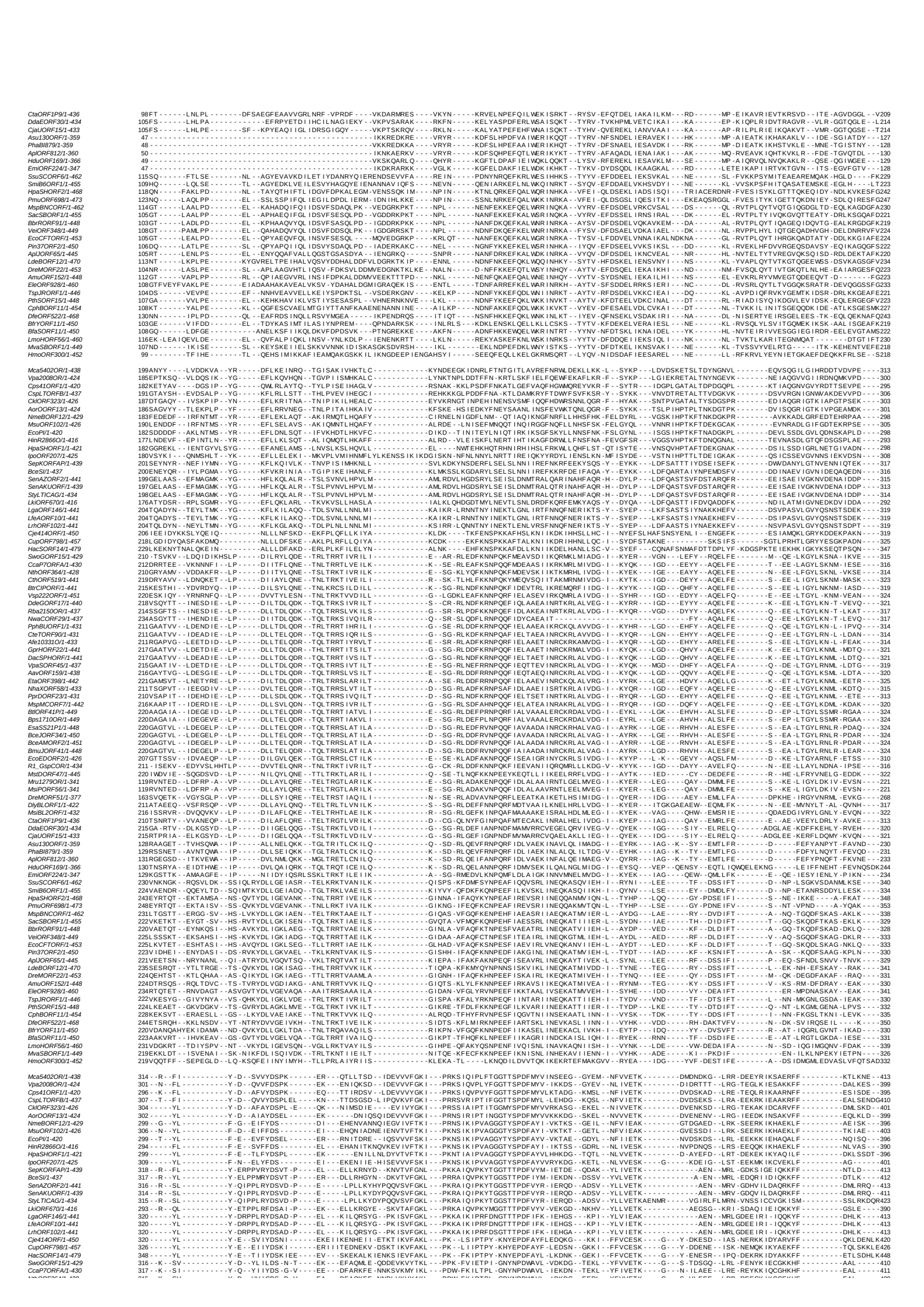


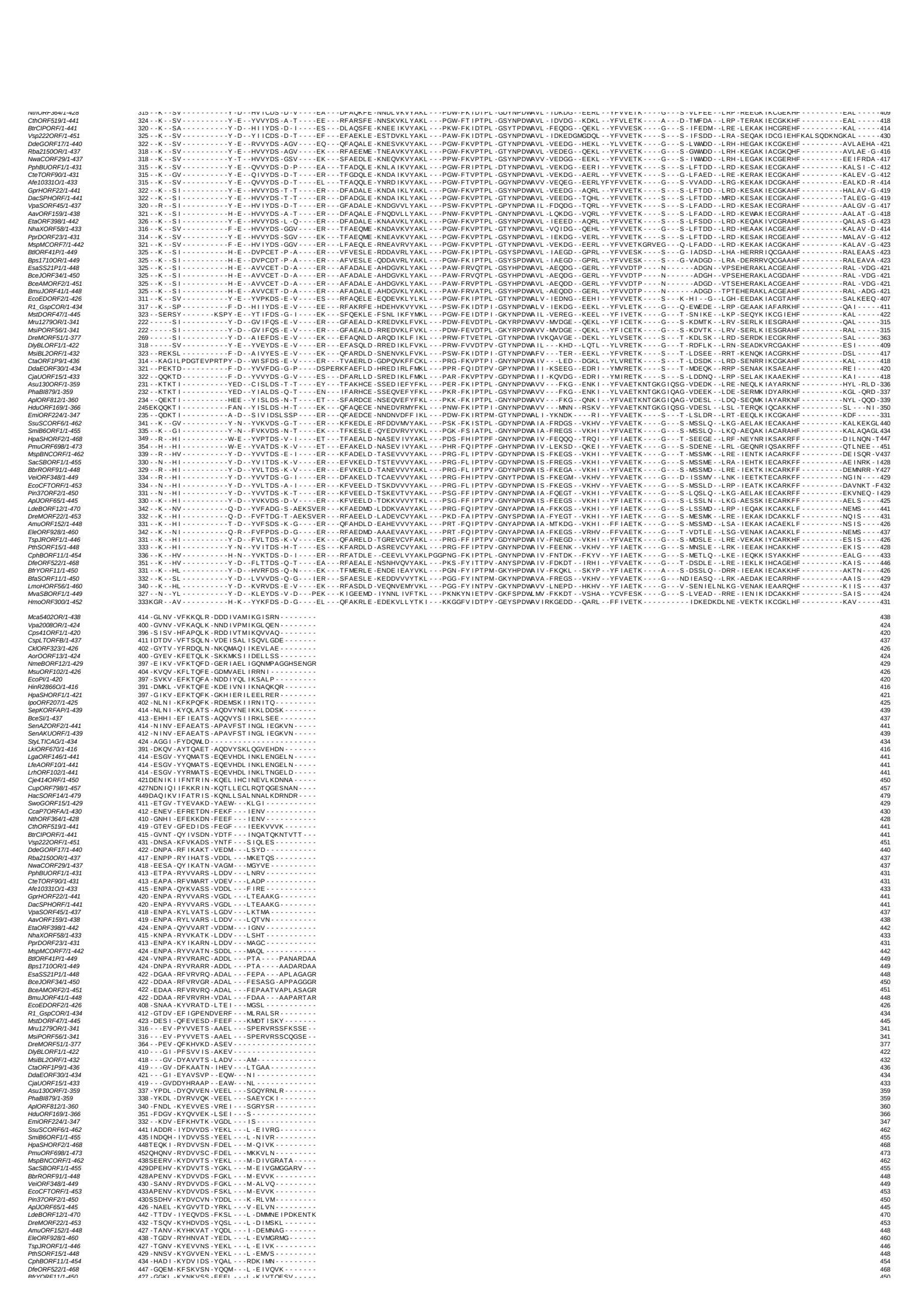




Type III nuclease







BfaSORF11/1-450	430 - T G E V - V Y S V V K T - Y Q D L - - - Y N A V T K - - - - - - - - -	450
LmoHORF56/1-460	438 - N S E V - T Y E V V D S - Y D K M - - - M D K L S S H I - - - - - - -	460
MvaSBORF1/1-449	425 - N E K I - I F D A V T S - V D N V - - - I D S I K P K A T V - - - - -	449
HmoORF300/1-452	432 - A Q D V - Q F K Q A K N - Y D D F - - - L L K I T V - - - - - - - - -	452

Table A1: Members of phylogenetic Group I

Group I		
ATPase	**Nuclease**	**Methylase**
CspLTORFB	Mca5402OR	EcoPI
Cps41ORF1	Vpa2008OR	EcoP15I
Mca5402OR	Cps41ORF1	MsuORF1026P
Vpa2008OR	CspLTORFB	HpaSHORF1827P
AorOORF13	CklORF323	HinR2866ORFAP
CklORF323	AorOORF13	IpoORF2074P
UbaNspORF	NmeBORF12	UbaNspORFAP
LgaORF146	MsuORF102	LgaORF1464P
LfeAORF10	HinR2866O	LrhGGORF981P
LrhORF102	EcoPI	LfeAORF1043P
SenAZORF2	EcoP15I	SenAZORF2565P
BceSI	HpaSHORF1	BceSI
LkiORF670	IpoORF207	LkiORF6700P
SepKORFAP	BceSI	Vpa2008ORF448P
EcoRP15I	SenAZORF2	Mca5402ORF234P
EcoPI	SenAKUORF	CspLTORFBP
NmeBORF12	StyLTI	Cps41ORF1462P
MsuORF102	LgaORF146	CklORF3239P
HinR2866O	LfeAORF10	AorOORF1317P
HpaSHORF1	LrhORF102	SepKORFBP
IpoORF207	LkiORF670	

Table A2: Members of phylogenetic Group II

Group II		
ATPase	**Nuclease**	**Methylase**
AmuORF152	DfeORF522	PmuORF698P
EleORF928	MvaSBORF1	HpaSHORF2158P
DreMORF22	BfrYORF11	AplJORF655P
LdeBORF12	BfaSORF11	TspJRORF1073P
PmuORF698	LmoHORF56	PthSORF151P
HpaSHORF2	SsuSCORF6	CphBORF1148P
BbrRORF91	SmiB6ORF1	BbrRORF912P
VeiORF348	HpaSHORF2	EcoCFTORF5372P
EcoCFTORF	PmuORF698	SacSBORF1825P
MspBNCORF	MspBNCORF	MspBNCORF543P
PcaORF272	SacSBORF1	LdeBORF1230P
SacSBORF1	BbrRORF91	AmuORF1521P
AplJORF65	VeiORF348	DfeORF5227P
Pin37ORF2	EcoCFTORF	MvaSBORF154P
TspJRORF1	Pin37ORF2	SsuSCORF608P
PthSORF15	PcaORF272	SmiB6ORF1249P
LmoHORF56	AplJORF65	BfaSORF1120P
DfeORF522	LdeBORF12	LmoHORF56P
SsuSCORF6	DreMORF22	EleORF802P
SmiB6ORF1	AmuORF152	TspX514ORF1828P
BfaSORF11	EleORF928	HmoORF3004P
BfrYORF11	TspJRORF1	BfrYORF1142P
CphBORF11	PthSORF15	
MvaSBORF1	CphBORF11	

Table A3: Members of phylogenetic Group III

Group III		
ATPase	**Nuclease**	**Methylase**
MhaPORFAP	Bgr4ORF41	VpaSORF4593P
PhaBI	SwoGORF15	MspMCORF787P
Asu130ORF	CcaP7ORFA	EtaORF3980P
AplORF812	NthORF364	EcoEDORF2295P
HduORF169	CthORF519	DacSPHORF4205P
EmiORF224	BtrCIPORF	GprHORF22920P
DdaEORF30	Vsp222ORF	Afe10331ORF1394P
CjaUORF15	R1.GspCOR	PprDORF237P
CtaORF1P	DdeGORF17	AavORF159P
MsiBL2ORF	Rba2150OR	BtlORF41P
DlyBLORF1	NwaCORF29	BglBGRORF160P
Mru1279OR	PphBUORF1	BceAMORF20P
MsiPORF56	CteTORF90	CteTORF908P
DreMORF51	Afe10331O	Rba2150ORF11466P
Bgr4ORF41	GprHORF22	NwaCORF2929P
R1.GspCOR	DacSPHORF	NhaXORF582P
SwoGORF15	VpaSORF45	NthORF364P
MstDORF47	AavORF159	SwoGORF1548P
CthORF519	EtaORF398	BtrCIPORF1929P
NthORF364	PprDORF23	GspCORF2154P
CcaP7ORFA	MspMCORF7	PhaBI
BtrCIPORF	BtlORF41P	Asu130ORF673P
Afe10331O	Bps1710OR	AplORF812P
CteTORF90	EsaSS21P	HduORF1691P
PphBUORF1	BceJORF34	EmiORF224P
Bps1710OR	BceAMORF2	Mru1279ORF580P
BtlORF41P	BmuJORF41	MsiPORF567P
BglBGRORF	BglBGRORF	MsiBL2ORF1946P
BceAMORF2	EcoEDORF2	DlyBLORF1612P
EtaORF398	Mru1279OR	DreMORF514P
AavORF159	MsiPORF56	CtaORF1P
VpaSORF45	DreMORF51	PcaPC1ORF848P
DacSPHORF	DlyBLORF1	DdaEORF3047P
GprHORF22	MsiBL2ORF	CjaUORF1580P
DdeGORF17	CtaORF1P	
Rba2150OR	DdaEORF30	
NwaCORF29	CjaUORF15	
PprDORF23	Asu130ORF	
MspMCORF7	PhaBI	
NhaXORF58	AplORF812	
EcoEDORF2	HduORF169	
	EmiORF224	
	MstDORF47	

APPENDIX II

Nucleic Acids Research Advance Access published March 14, 2016

Nucleic Acids Research, 2016 **1**
doi: 10.1093/nar/gkw154

Structural insights into DNA sequence recognition by Type ISP restriction-modification enzymes

Manasi Kulkarni[1,†], Neha Nirwan[1,†], Kara van Aelst[2], Mark D. Szczelkun[2] and Kayarat Saikrishnan[1,*]

[1]Division of Biology, Indian Institute of Science Education and Research, Pune 411008, India and [2]DNA-Protein Interactions Unit, School of Biochemistry, Medical Sciences Building, University of Bristol, Bristol BS8 1TD, UK

Received December 21, 2015; Revised February 26, 2016; Accepted February 29, 2016

ABSTRACT

Engineering restriction enzymes with new sequence specificity has been an unaccomplished challenge, presumably because of the complexity of target recognition. Here we report detailed analyses of target recognition by Type ISP restriction-modification enzymes. We determined the structure of the Type ISP enzyme LlaGI bound to its target and compared it with the previously reported structure of a close homologue that binds to a distinct target, LlaBIII. The comparison revealed that, although the two enzymes use almost a similar set of structural elements for target recognition, the residues that read the bases vary. Change in specificity resulted not only from appropriate substitution of amino acids that contacted the bases but also from new contacts made by positionally distinct residues directly or through a water bridge. Sequence analyses of 552 Type ISP enzymes showed that the structural elements involved in target recognition of LlaGI and LlaBIII were structurally well-conserved but sequentially less-conserved. In addition, the residue positions within these structural elements were under strong evolutionary constraint, highlighting the functional importance of these regions. The comparative study helped decipher a partial consensus code for target recognition by Type ISP enzymes.

INTRODUCTION

Protein–nucleic acid interactions are central to a large number of important cellular functions. The interactions can be broadly classified into either sequence specific or sequence independent interaction. Sequence independent binding is primarily established by protein residues via ionic and hydrogen bonds with the sugar-phosphate DNA backbone, and/or by stacking interactions with bases. Sequence specificity is established via base-specific interaction made by protein residues (1). Understanding the molecular details of how proteins bind to nucleic acid has not only contributed to the understanding of the biology of the system, but has also contributed to our ability to engineer proteins with new binding specificity.

Engineering proteins to bind specific nucleic acid sequences has been in the forefront of the research activities driving biotechnological breakthroughs. Successful examples of reagents generated from nucleic acid binding proteins include zinc-finger domains coupled to an endonuclease, where the zinc-fingers have been engineered to recognize different target sites, and transcription activator-like effector nucleases (TALENs), made of nuclease-fused arrays of TAL domains that recognize specific sequences (2). Furthermore, Pumilio FBF homology (3) and pentatricopeptide proteins (4) have been engineered for new RNA sequence specificity. However, engineering new specificities has been hitherto successful using repeating modular and non-catalytic units that recognize short nucleic acid sequences (a single or a few nucleobases).

Engineering specificities of other nucleic acid binding proteins has in general had poor success. This is despite detailed understanding of how a large variety of proteins that participate in different cellular functions recognize specific DNA sequences. A limited exception to this is the engineering of new specificities in meganucleases (5). Among DNA binding proteins, restriction enzymes were amongst the first that were studied biochemically, biophysically and structurally towards rationally designing new specificities. For example, detailed studies of the enzymes BamHI, BglII and BstYI, which recognize similar sequences, demonstrated that the enzymes use different recognition strategies, which hindered rational design of new specificities (6–8).

Subsequent efforts included use of the MmeI family of Type IIL restriction-modification (RM) enzymes, which have both endonuclease and methyltransferase activities within the same polypeptide. This offers the advantage that change in specificity is produced in both the destructive nu-

*To whom correspondence should be addressed. Tel: +91 2025908047; Fax: +91 2025908186; Email: saikrishnan@iiserpune.ac.in
†These authors contributed equally to the paper as first authors.

clease and protective methyltransferase. Based on sequence analyses of known Type IIL enzymes and their target sequences, mutations were carried out to generate changes at three positions of their six base pair sequences (9). However, change in specificity at other positions was less successful (9,10). Here, we report the structure of the single polypeptide Type ISP RM enzyme LlaGI, allowing a structural comparison of DNA recognition elements with the homologous Type ISP enzyme LlaBIII (11).

Type ISP enzymes are similar in domain organization to Type IIL enzymes, except for the insertion of a helicase-like ATPase domain motor that plays an essential role in nucleolytic activity (Figures 1 and 2A). RM enzymes prevent the integration, replication and expression of foreign DNA in the host bacterium by introducing double-strand (ds) DNA breaks into the invading genomes. In case of Type ISP enzymes, the entire gamut of tasks is carried out by the coordinated action of target recognition, methyltransferase (MTase), helicase-like ATPase and nuclease domains that constitute the single polypeptide chain (12). The target recognition domain (TRD) recognizes a specific DNA sequence and the methyltransferase modifies the target adenine. Nucleolytic cleavage of unmodified DNA is triggered by recognition of two DNA binding sites on the same DNA, ATP-dependent long-range communication, and collision of two RM enzymes (Figure 1) (13,14). LlaBIII and LlaGI are the two prototypes of Type ISP RM enzymes (15,16). LlaGI recognizes 5′-CTnGAyG-3′, while LlaBIII recognizes 5′-TnAGCC-3′ (Figure 1) (15,16). DNA cleavage requires the presence of two sites on the same DNA, in an inverted, head-to-head repeat as defined by the arrowheads in Figure 1. Cleavage occurs at a distant location between the two sites, with the distribution of locations concentrated on the midpoint position (11,17).

We recently reported a high-resolution crystal structure of LlaBIII (11). The structure of LlaBIII revealed the molecular architecture and domain organization in a Type ISP enzyme; the mode of target recognition; a mechanism for coupling target recognition and ATPase domain activation; a unique mechanism for translocation coupled nucleolytic activity (11). We found that the LlaBIII nuclease domain was located upstream of the DNA target and ATPase domain. This architecture was inconsistent with the model proposed for DNA cleavage (13), which, akin to Type I enzymes (Figure 1), involved translocation-mediated DNA looping resulting in convergence and collision of two enzymes leading to DNA cleavage. Based on complementary single-molecule magnetic tweezers assay and single cleavage sequence analysis, we found that Type ISP enzymes translocated along DNA without looping, and proposed a nucleolytic mechanism in which DNA break formation resulted from multiple nicks caused by distal nucleases in the collision complex of the enzyme (Figure 1) (11,18).

In continuation of our study of the mechanism of action Type ISP enzymes, we report here the co-crystal structures of a nuclease mutant of LlaGI and a nuclease-deleted mutant LlaGIΔN, bound to DNA substrate mimics. As mentioned earlier, the target of LlaGI is distinct from LlaBIII. But the two enzymes share an amino acid sequence identity of ∼80%, with the target binding MTase-TRD unit having an identity of 58% (Figure 2A). The structures of LlaGI and LlaBIII provided us with a unique opportunity to understand and compare in atomic detail how the two closely related RM enzymes recognized disparate target sequences and characterize the recognition region. Similar studies helped understand the complexities of DNA sequence recognition by zinc-finger proteins and served as a platform in engineering novel zinc-fingers (19).

We also analysed 552 amino acid sequences of Type ISP enzymes to study the conservation of the target recognition region in this family of enzymes. Amino acid sequence comparison of 11 Type ISP enzymes whose targets are known obtained from the REBASE database (20) provided further insights into the correlation between target sequence and amino acids involved in their recognition. This structure comparison and sequence analyses show that target recognition by Type ISP enzymes is complex and that change in target recognition cannot be achieved only by corresponding change in the contacting amino acids. Changes at additional positions are often required to generate specific interactions with the new base. Though a simple code for target recognition could not be obtained, the study led to a consensus and a predictive code that may facilitate engineering new specificities.

MATERIALS AND METHODS

Purification of LlaGI and LlaGIΔN

Native 180 kDa LlaGI was overexpressed in 10L of *Escherichia coli* BL21 (DE3) from a recombinant clone of the *llagi* gene in the pRSF vector (15). The highest amount of soluble protein in the crude lysate was observed upon inducing the culture at 25°C with 0.5 mM IPTG at $OD_{600} = 0.6$. The incubation temperature was lowered from 37 to 25°C before addition of IPTG. Induced cells were harvested after 5 h further incubation. To avoid proteolytic degradation of LlaGI, protease inhibitors (Roche, UK) were added to the lysis buffer (50 mM Tris-HCl pH 8.0, 150 mM NaCl, 10 mM $MgCl_2$, 1 mM EDTA and 1 mM DTT), all steps of purification were carried out at 4°C, and the protein was purified to homogeneity within 24 h of lysis. The cells were lysed by sonication. LlaGI was salted-out using 70% w/v ammonium sulphate. The re-suspended protein pellet was further purified by column chromatography using heparin followed by MonoQ. The strongly anionic MonoQ column not only removed protein impurities but also any cellular DNA bound to the enzyme. Finally, size exclusion chromatography using Superdex 200 10/300 (GE Healthcare) ensured homogenous monomeric LlaGI. Equivalent strategies were followed for purification of LlaGIΔN (11). Purified LlaGI and LlaGIΔN were stored in a buffer containing 10 mM Tris-HCl pH 7.4, 100 mM NaCl and 1 mM DTT. Purified samples stored at −80°C remained intact and could be thawed and used for crystallization later.

Purification of DNA for crystallization

The duplex DNA substrate mimics used for co-crystallization were obtained by annealing the oligos 5′ -TTAGCTAATAGACTGGATGGAGG-3′ and 5′-TCCT CCATCCAGTCTATTAGCTA-3′ for LlaGIΔN-DNA, and 5′-GCTCTAGCTAATAGACTGGATGGAGGTG-3′

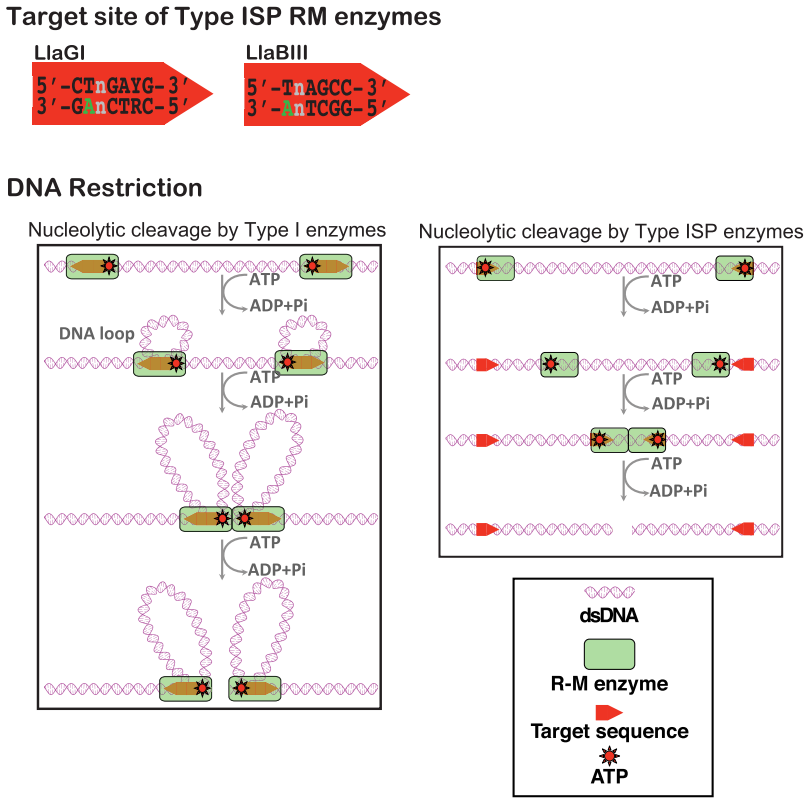


Figure 1. Cartoon illustrating the target sequence of LlaGI and LlaBIII and the translocation-coupled nucleolytic cleavage of DNA by Type I and Type ISP enzymes. Type I enzyme complexes have two MTase, ATPase and nuclease domains, and hence can translocate and cleave DNA either upstream or downstream of the target. For clarity, translocation upstream is not illustrated. Note that in the schematic of Type ISP enzymes ATP is located at the far end to indicate that the ATPase domain is upstream of the target.

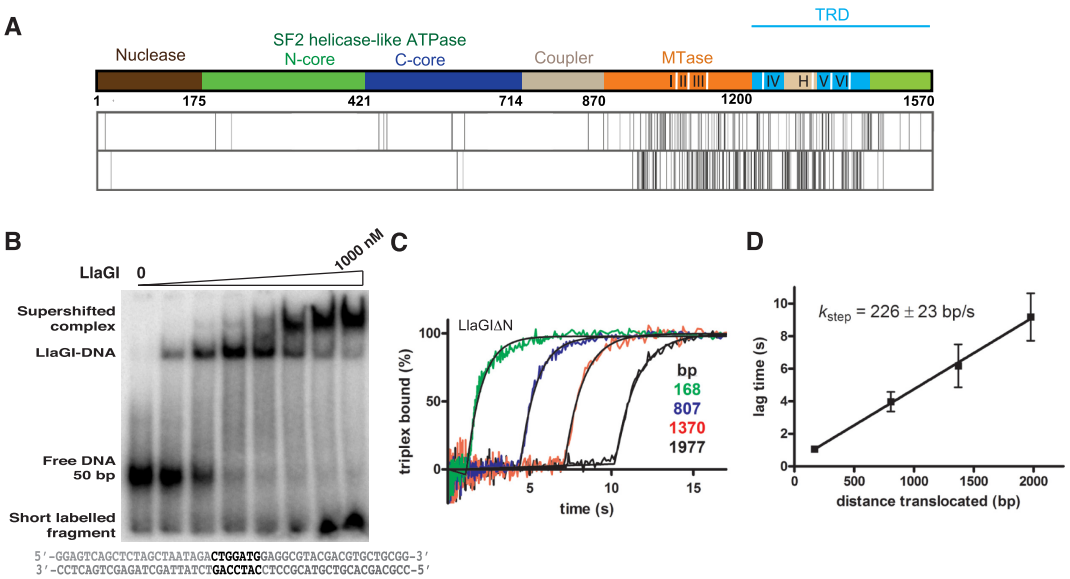


Figure 2. (**A**) The domain organization of LlaGI. The positions of the structural elements that recognize the targets are shown as white lines. The amino acid sequence of LlaGI and LlaBIII were aligned pairwise to identify regions of homology, similarity and difference (4). The vertical lines show positions that have amino acid substitutions that are similar (upper row) or different (lower row)—all other positions are identical. (**B**) EMSA assays of LlaGI binding to 10 nM of a 50 bp DNA substrate with 0, 10, 25, 50, 100, 250, 500 and 1000 nM protein. (**C**) Triplex displacement reactions on linear DNA. Triplex DNA (with different spacing between the LlaGI site and fluorescent triplex) were pre-incubated with enzyme and the reactions initiated with ATP, to give a final concentration of 1 nM DNA, 100 nM LlaGIΔN and 4 mM ATP at 25°C. The triplex displacement profiles, which have lag phases characteristic of a translocating motor protein, were fitted to Equation (1) to obtain the lag time (T_{app}). (**D**) The linear relationship between T_{app} and d was used to estimate the translocation rate. The points are the mean and SD for repeat reactions measured using two different preparations of LlaGIΔN.

and 5′ -CACCTCCATCCAGTCTATTAGCTAGAGC-3′ for LlaGI-DNA complex. The chemically synthesized and PAGE-purified oligos were purchased from Integrated DNA Technologies, USA. The single-strand DNA were annealed and the duplex DNA was purified using a MonoQ column. The pure duplex DNA was concentrated and stored in sterile water at −20°C.

Electrophoretic mobility shift assay

7% w/v polyacrylamide gels were used to study binding of LlaGI to DNA by electrophoretic mobility shift assay (EMSA). Duplex DNA was generated from complementary single-strand DNA as described above. The 5′-ends of the duplex DNA substrates were labelled with ^{32}P using T4 polynucleotide kinase from New England Biolabs, USA. The binding reaction buffer contained 50 mM Tris-HCl pH 7.4, 100 mM NaCl, 10 mM $MgCl_2$, 0.01 mg/ml BSA, 1 mM DTT and 10% (v/v) glycerol. The reactions were incubated at 4°C, and the DNA–protein complexes resolved by electrophoresis at room temperature. The gels were dried using a gel dryer, exposed to a phosphor screen, and images recorded using a Typhoon Imager (GE healthcare).

Translocase assay

Triplex displacement measurements were carried out in an SF61-DX2 stopped-flow fluorimeter as described previously (16). The sample temperature was maintained at 25°C by a water bath connected to the chamber housing the syringes and flow cell. Reactions were started by adding ATP to a DNA/enzyme mix with the final reaction conditions as 1 nM linear DNA (0.5 nM tetramethylrhodamine triplex), 100 nM LlaGIΔN and 4 mM ATP in 50 mM Tris-HCl, pH 8.0, 10 mM $MgCl_2$, 1 mM DTT. Data averages were collected from at least three individual time-courses and analysed in KinetAsyst 3.11 (Hi-Tech Scientific) and Prism 4 (GraphPad software, Inc., San Diego). Triplex displacement data were fitted to Equation (1) (modified from Gilhooly and Dillingham, 2014) which defines displacement (Y) as a function of time (t), as the sum of a background linear rate (with gradient m) and an offset exponential with an x-axis offset at times T_{app} (the apparent 'lag time'), with amplitude A, an apparent rate constant of triplex displacement of k, and an offset ($C1$) which is the value for free triplex at $t = T_{app}$:

$$Y = (x < T_{app})(mx) + (x > T_{app})\left(\left(A\left(1 - \left(\exp^{-k(t-T_{app})}\right)\right)\right) + C1\right) \quad (1)$$

The variation of T_{app} (s) with distance (d, bp) was fitted to:

$$Y = \left(\frac{1}{k_{step}}\right) d + C2 \quad (2)$$

Where k_{step} is the rate of translocation and $C2$ is the intercept with the y-axis. Although C2 can be used to estimate the initiation time, there are limitations in doing this (16), so this value is not reported here. Note, this uncertainty has no effect on the quality of the k_{step} value.

Crystallization and data collection

A complex of the purified protein and DNA substrate was formed by mixing the two in 1:1.3 molar ratio at 4°C. A protein concentration of 5 mg/ml and a crystallization drop size of 200 nl were used for all the initial crystallization trials by sitting drop vapour diffusion method. The nanodrops were set using a robotic liquid handler. Over 4000 different conditions of varying buffers, additives, precipitants, DNA substrates and temperatures were screened for crystallization of LlaGI-DNA and LlaGIΔN-DNA complex. The most promising conditions were further optimized for growing single crystals using an optimization grid. Crystals grown in nanodrops did not diffract well. Improvement in the diffraction quality of LlaGI-DNA crystals was achieved from larger crystals grown in 2 μl drops at 291 K by the sitting drop method. One such crystal of LlaGI-DNA grown from a 1:1 ratio of protein:reservoir buffer (100 mM MES pH 5.6, 300 mM KCl, 6.5% w/v PEG 20,000) diffracted to 7.40 Å at 100 K, with glycerol as a cryoprotectant. Suitable diffracting crystals of LlaGIΔN were grown in 4 μl hanging drops at 291 K by the vapour diffusion method. The reservoir buffer contained 100 mM Tris-HCl pH 7.4, 20% w/v PEG 20 000, 4% w/v PEG 550 MME and 150–250 mM sodium acetate. One of these crystals diffracted to 2.84 Å at 100 K, with ethylene glycol as cryoprotectant.

Structure determination

The LlaGI-DNA and LlaGIΔN-DNA diffraction data were collected at the European Synchrotron Radiation Facility ID23–1 and the Diamond Light Source I03 beamlines, and processed using MOSFLM (21) and XDS (22), respectively. The intensities were scaled and merged using AIMLESS (23). The structure solution for LlaGIΔN-DNA crystal was obtained by molecular replacement using the program PHASER (24) and the structure of partially built LlaGIΔN^{Se}-DNA (11) as the search model. The coordinates of the ATPase domain of LlaBIII (PDB ID: 4XQK) were used as a guide for building the ATPase domain. The structure solution for LlaGI-DNA crystal was obtained by molecular replacement using the structures of coupler-MTase-TRD unit of LlaGIΔN and nuclease-ATPase unit of LlaBIII (PDB ID: 4XQK) as separate search models.

Initial cycles of structure refinements were carried out by REFMAC5 (25) and subsequently by *phenix.refine* (26). The maps were visualized and model building carried out using COOT (27). In case of LlaGIΔN-DNA structure, positional and isotropic B-factor refinement was carried out. Due to the low resolution of LlaGI-DNA crystal data, only a domain-wise rigid-body refinement was carried out keeping the B-factors constant, and side chain atoms beyond Cβ were truncated and not included in the refinement.

Amino acid sequence analysis

A database of amino acid sequences of 552 Type ISP enzymes were generated by BLAST (28) using the LlaGI sequence and the non-redundant protein sequences database. Multiple sequences with 100% conservation were reduced to a single sequence. Many of the sequences showed very high sequence conservation in the first half of the ORF. However, much higher levels of variation were observed in the C-terminal half of the MTase domain and in the TRD. Therefore the sequences were not further reduced for redundancy. The sequences were aligned using Clustal Omega

(29) and Jalview (30) (Supplementary Data 1). The evolutionary conservation scores for the residues in the MTase-TRD unit were calculated using ConSurf (31). To plot, the scores obtained from ConSurf were multiplied by −1 and scaled up by addition of a constant value of 3.505 so that the lowest conservation score was zero. Evolutionary constraint (EC) strengths for the residues were calculated using the program EVcouplings available online (32). As the program could only handle a sequence length of 600 residues, residues starting from LlaGI-868 to LlaGI-1428 were used for the calculations. Recognition sequences for a subset of the Type ISP enzymes were obtained from REBASE (20) and are based on DNA methylation patterns determined from PacBio sequencing.

RESULTS

A nuclease-dead N-terminal deletion mutant of LlaGI is an active translocase

During characterization of LlaGI-DNA complex by EMSA prior to co-crystallization studies, a nucleotide-independent cleavage of single-site substrates resulted in release of short-labelled fragment (Figure 2B). This suggested that even in the absence of ATP, wild-type LlaGI carried a residual nucleolytic activity. To protect the integrity of the DNA during crystallization of LlaGI-DNA complexes, two different nuclease-dead mutants of LlaGI were generated. One of these was a previously characterized point mutant in which the catalytic aspartate at position 78 was mutated to alanine (13). This mutant has wild-type translocation activity. In addition, we sought to generate truncations of LlaGI that removed the nuclease activity whilst retaining translocase activity. Identification of suitable exposed inter-domain loops using partial proteolytic digestion proved unsuccessful (data not shown). As an alternative, we generated a secondary structure prediction of LlaGI using the PSIPRED server (Supplementary Figure S2) (33). From this the region 160–180 was predicted to be a loop between the nuclease and ATPase domains. We chose to delete the residues 2–165, producing the recombinant protein LlaGIΔN (11).

Purification of LlaGIΔN was achieved using the same protocol as with wild-type LlaGI, and the protein did not show any differences in solubility/stability during the preparation. To check for the retention of translocase activity, we utilized the triplex displacement assay (Figure 2B). The displacement profiles show the distance-dependent lag time (T_{app}) characteristic of a translocating motor and we were able to measure a stepping rate of 226 ± 23 bp/s at 25°C (Figure 2C), in good agreement with the value for the wild-type enzyme under these conditions (250 ± 10 bp/s at 25°C) (17). Moreover, LlaGIΔN was also able to cooperate with Type ISP enzymes to activate DNA nicking at distant non-specific sites (17), consistent with DNA translocation. Both the D78A and ΔN mutants were used for crystallographic studies.

Molecular architecture of LlaGI bound to DNA

The structures of LlaGIΔN bound to a 22-bp DNA and full-length LlaGI bound to a 28-bp DNA were determined to a resolution of 2.84 and 7.4 Å, respectively (Table 1). In the LlaGIΔN-DNA crystals there were two molecules in the asymmetric unit, while in LlaGI-DNA there were four. Like LlaBIII, LlaGI contained six structural domains: the N-terminal Mrr-family nuclease, followed by the N-core and C-core RecA folds of the SF2 helicase-like ATPase domain, connected to the γ-class of N6-adenine MTase domain and the C-terminal TRD by an all α-helical coupler domain (Figure 3A and B). Comparison of LlaGI-DNA, LlaGIΔN-DNA and LlaBIII-DNA structures revealed interdomain conformational mobility (Figure 3C).

Like in the case of the LlaBIII-DNA complex (11), in both LlaGI-DNA and LlaGIΔN-DNA structures, the MTase domain and TRD held the target in a vice-like grip. The T:A base pair of the LlaGI target that is methylated was designated as position +1 based on the convention used by Chand *et al.* (11) to describe the LlaBIII target. Therefore the first C:G base pair of the LlaGI target was designated as −1. In LlaGIΔN-DNA, the upstream end of the oligonucleotide was not long enough to interact with the ATPase domain (Figures 3B and 4). Instead, this end interacted with the ATPase domain of the neighbouring molecule of the asymmetric unit. In the LlaGI-DNA structure, the base pairs at the ends of the DNA could not be built because of poor density. In the LlaGIΔN structure, only a part of the C-core of chain B could be built with confidence, while the C-core of chain A could not be built due to poor electron density. The lack of electron density could be the result of conformational flexibility of the domains in the crystal lattice.

The structures of two close homologues, LlaGI and LlaBIII, and those of other evolutionarily related multidomain restriction enzymes (Type I and Type IIG), provided us with an opportunity to carry out a detailed comparison of their structure to understand target recognition. This is detailed in the following sections. All the residue numbering in the following sections are based on LlaGI sequence unless specified otherwise.

The target binding MTase-TRD unit

The structures of LlaGI and LlaBIII showed a conserved mode of DNA binding at the target by a clamp formed by the MTase domain and TRD. The structures of the TRDs of LlaGI and LlaBIII are very similar (RMSD = 1 Å) despite the relatively low amino acid sequence identity of ∼52% (Figure 2A). Structural variation in the form of insertion or deletion of amino acid residues are primarily located in regions not directly involved in DNA recognition (Supplementary Figure S3A). A similar mode of target binding is employed by the prototypical γ-class of N6-adenine MTase domain M. *Taq*I (34). However, the TRD of the Type ISP enzymes makes much more extensive interactions with the DNA (Figure 4) (11). Additionally, the TRD and MTase domain of the Type ISP enzymes form a closed clamp around the target, while in M. *Taq*I the two domains form an open DNA binding cleft. The TRD of LlaGI is larger than that of M.*Taq*I by ∼200 residues and can be divided into three structural subdomains—the core (1200–1239 and 1297–1440), the jaw (1240–1296) and the guide (1440–1578) as illustrated in Figure 5A. Similarly, LlaBIII

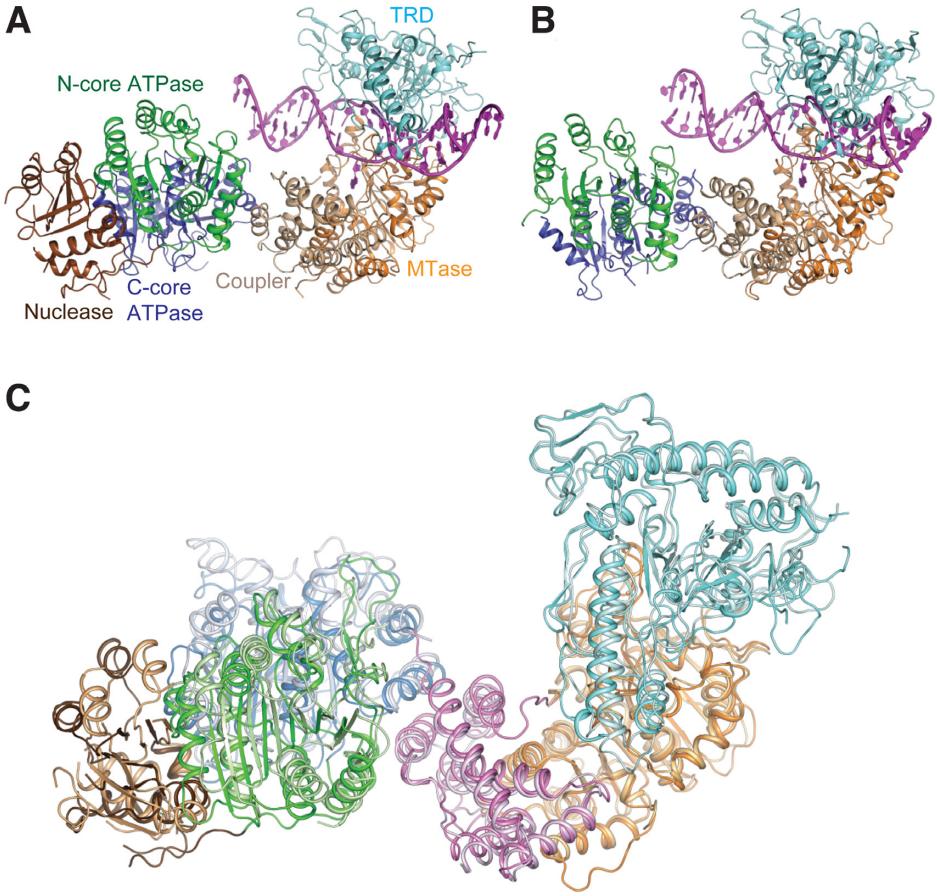


Figure 3. Ribbon representation of (**A**) LlaGI-DNA complex at 7.4 Å resolution and (**B**) LlaGIΔN-DNA complex at 2.8 Å resolution. The six structural domains are coloured distinctly. The nuclease is coloured brown, the N-core of the ATPase domain is coloured green, the C-core is in blue, the coupler in beige, the MTase domain in orange and the TRD in cyan. (**C**) The superposition of the Cα-trace of LlaGI at 7.4 Å and LlaBIII (PDB ID: 4XQK) reveals domain movement. LlaGI was superposed on to LlaBIII with respect to the coupler using Coot (27).

Table 1. Data collection and refinement statistics

	LlaGI-DNA	LlaGIΔN-DNA
Data collection		
Space group	I2	$P2_1$
Cell dimensions		
a, *b*, *c* (Å)	267.9,203.4,291.5	87.4,222.3,117.4
α, β,γ (°)	90.0,96.2,90.0	90.0,105.1,90.0
Resolution (Å)	50.0–7.4 (8.11–7.40)	50.0–2.84 (2.99–2.84)
R_{sym}	8.0 (50.7)	8.9 (58.8)
I / σ(I)	8.0 (2.0)	9.6 (2.0)
Completeness (%)	96.7 (98.4)	99.8 (99.9)
Redundancy	2.0	2.0
Refinement		
Resolution (Å)	50–7.4	50–2.84
No. of reflections (total/test)	20176/1040	101504/5107
R_{work}/R_{free}	34.6/37.4	22.9/26.2
R.m.s. deviations		
Bond lengths (Å)	0.015	0.004
Bond angles (°)	1.281	0.767
Ramachandran plot (%)		
Favoured	95.4	95.4
Outliers	0.2	0.2

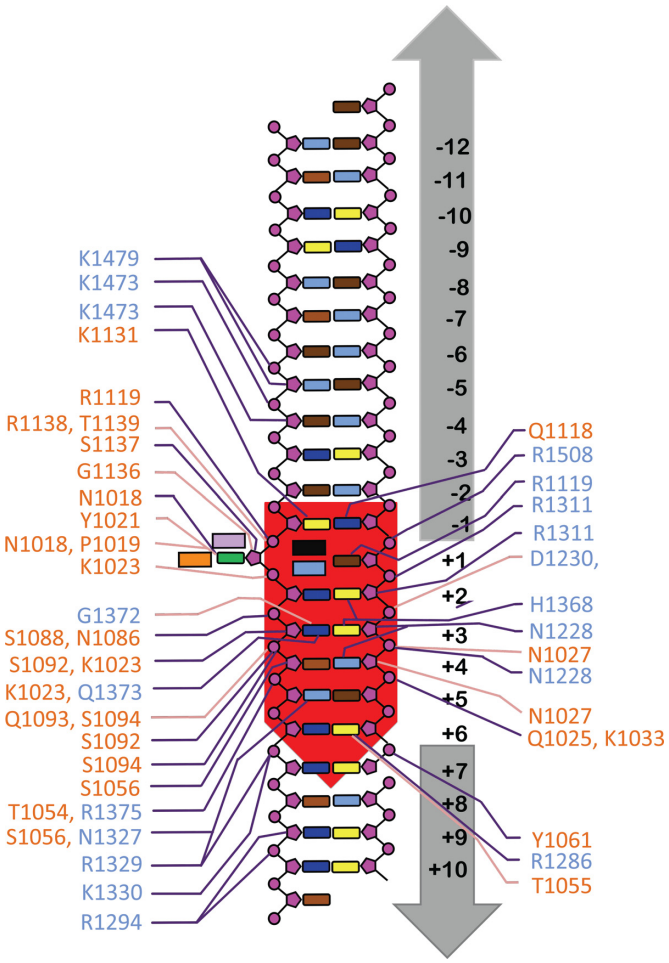


Figure 4. A schematic diagram of protein–DNA interactions in LlaGIΔN-DNA complex. Amino acid colours are according to domains in Figure 2A. DNA bases are cyan (A), brown (T), blue (C) and yellow (G), main chain interactions are pink lines, side chain interactions are purple lines. The target adenine (green) is +1, with downstream positions defined as positive and upstream positions defined as negative. Residues Y1021, F1133, R1119 and R1138 are depicted as lavender, orange, sky blue and black coloured boxes, respectively.

TRD (Figure 5B) can be delineated into the core (1205–1244) and (1291–1448), the jaw (1245–1288) and the guide (1448–1578).

An analysis of a multiple sequence alignment (MSA) of 552 Type ISP enzyme sequences across the target binding MTase-TRD unit showed that the MTase domain was more strongly conserved in comparison to the TRD (Figure 5B and C). Amongst the TRD subdomains, the jaw appeared to be the least conserved, followed by the core, while the guide appeared to be the most conserved. Structurally, the core has the same fold as the TRD of M.*Taq*I (Supplementary Figure S3B), TRD1 or TRD2 of Type I enzymes (Figure 5B) (35) and the TRD of the Type IIG enzyme BpuSI (Supplementary Figure S3C) (36). This subdomain along with the MTase domain is the primary reader of the target sequence (see below). Visual inspection of the amino acid conservation score mapped onto MTase-TRD unit suggested that the surface of the TRD core that contacted the DNA was not strongly conserved, unlike the recognition elements of the MTase domain (Figure 5B and C). Amino acids of the central β-sheet of the TRD also showed moderate conservation. However, the helices (I1395–I1428) stacked on the central β-sheet were relatively well conserved.

The jaw, an insertion in the core, is made of a three-helix bundle (Figure 5A), which, based on structural similarity search using DALI server (37), was not found in other RM enzymes. The sequence alignment indicated the amino acids of this subdomain to be the least conserved. LlaGI also has inserts in this region compared to LlaBIII (Supplementary Figure S3A). The three helices of the jaw along with the MTase loop S1022-I1035 seal the clamp around the DNA (Figure 5D). One of the jaw helices (1286–1295) fits into the major grove of the deformed target and interacts with the phosphate backbone at +9 (a non-specific position). The clamp may aid the processivity of the translocating enzyme.

The structure of the closed clamp bound to the DNA raises the question as to how the enzyme would assemble on its DNA substrates, which could be a long linear or a closed circular DNA. Comparison of the Type ISP-DNA structures with the apo-structure of BpuSI (36) provided an insight into this question. Superposition of the MTase domain of LlaGI and BpuSI revealed that the TRDs of the two enzymes were rotated by 92° with respect to one another. Consequently, the TRD in the apo-structure of BpuSI would be predicted to be in an open conformation to allow entry of DNA. We suggest that in the apo-form of the Type ISP enzymes, the TRD-MTase interdomain conformation would be in a similar open conformation, facilitating the DNA to slide into a cleft created by the two domains (Figure 5D).

With the limited structural information on Type ISP enzymes, it is difficult to postulate on the transition from the open to the closed conformation. However, deriving inspiration from the studies on the crystal structures of the Type II enzyme BamHI in its apo form, bound to non-specific DNA and specific DNA (38), it may be envisaged that MTase-TRD unit of Type ISP enzyme would transit from the open conformation when not bound to DNA, to a conformationally distinct, and as yet unobserved, state when bound to non-specific DNA, to the closed conformation seen in the crystal structure when bound to specific DNA. The trigger for these conformational changes could be the interactions made by the protein with non-specific and specific DNA. We had earlier postulated that translocation initiation would be accompanied by remodelling of the target bound Type ISP enzyme (11,18).

The C-terminus of the core continues into the guide. The guide has a two layer open-faced β-sandwich structure with a four-stranded N-type Greek key motif layer stacked against a layer of three helices with a β-hairpin loop inserted between two of them (Figure 5A). Based on a Dali search, the fold of the guide was not found in other RM enzymes. Amongst the three subdomains, amino acids of the guide were found to have the highest conservation scores (Figure 5B). The helices at the N-terminus and C-terminus of the subdomain form a bundle, reminiscent of the helical bundle formed by the Central and Distal Conserved Regions of the HsdS subunit of classical Type I enzymes (35,39).

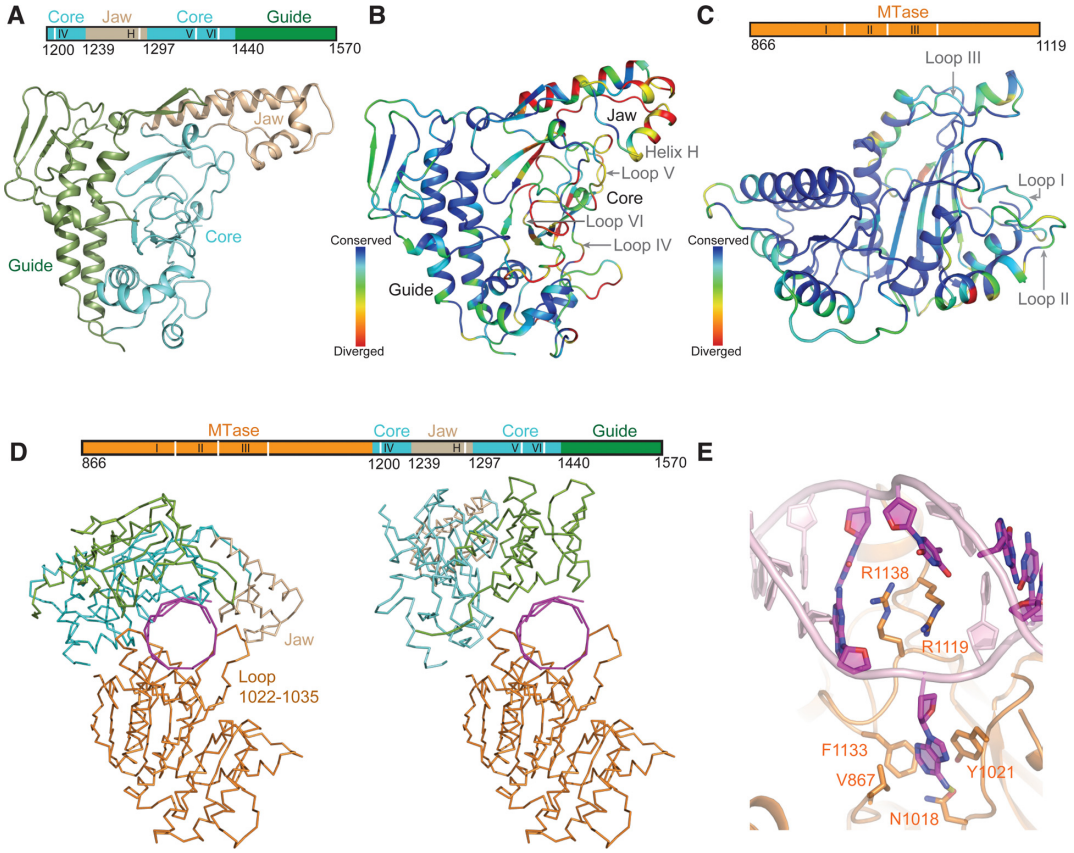


Figure 5. TRD of Type ISP enzyme. (**A**) The three subdomains seen in TRD of LlaGIΔN: the core is coloured cyan, the jaw is coloured wheat and the guide is coloured green. The positions of the structural elements that recognize the targets are shown as white lines in the cartoon above the ribbon diagram. (**B**) Mapping of the evolutionary conservation scores of the Type ISP TRD residues onto the LlaBIII-TRD and (**C**) onto the LlaBIII-MTase domain structure. The positions of the structural elements that recognize the targets are shown as white lines in the cartoon above the ribbon diagram. (**D**) Left: the closed structure of the clamp formed by MTase domain and TRD in LlaGIΔN around the DNA (magenta). Right: the open structure of the clamp modelled based on the apo structure of BpuSI (PDB ID: 3S1S) (36). (**E**) The active site of the LlaGIΔN MTase domain with the target adenine flipped. The structural elements of the MTase domain and TRD that are involved in target recognition are marked.

In Type I enzymes, the helical bundle acts as a molecular ruler, spacing the two TRDs for recognition of the two halves of the bipartite target sequences (e.g. GAANNNNNNRTCG for EcoR124I). In the Type ISP enzymes, the helical bundle together with the hairpin loop insertion in the guide forms a concave surface against which the core is packed (Figure 5A). The guide does not participate in target recognition. Instead, the Greek-key motif located at the end of the helical bundle and the hairpin loop steers the DNA upstream of the target. K1473, K1479 and R1508 of the LlaGI guide interact with the phosphate backbone (Figure 4). As in case of LlaBIII (11), the DNA is steered towards the ATPase domain.

The MTase domain is structurally similar to other γ-class adenine-N6 MTases, including the prototypical M. *Taq*I, and the MTase domains of Type I and the Type IIG protein BpuSI. As in the case of LlaBIII, the target adenine to be methylated is flipped out via the minor groove into the active site of the LlaGI MTase domain despite the absence of AdoMet (Figure 5E). The interactions made by the flipped adenine are similar to that made with AdoMet-bound M. *Taq*I. The adenine stacks against Y1021 and makes an edge-to-surface T-interaction with the aromatic F1133. The adenine-N6 is hydrogen bonded to the catalytic N1018. The cavity formed by the base flip is filled by the intercalation of two residues R1119 and R1138 (Figure 5E). The two residues along with the flanking bases at positions +2 and −1 form a continuous stack. Intercalation by two positionally identical residues R1119 and M1137 fills the cavity in the LlaBIII-DNA complex.

The use of two residues to fill the cavity arising from the extrahelical base is unique to Type ISP enzymes. In case of the prototype M. *Taq*I, domain movement compresses the DNA locally (34) and P393 partially intercalates at the position of base flip. In most other enzymes that employ base flipping to catalyse DNA modification, such as other classes of DNA methyltransferases and DNA repair enzymes (40), the cavity is filled in by a single, often hydrophobic residue. However, the cavity formed upon base flipping by Type ISP enzymes appears larger than that seen in case of other enzymes due to the large distortion in the structure (Supplementary Figure S4). Like the DNA bound to LlaBIII (11), the DNA is bent by ∼34° at the target. The width of both the major and minor grooves at the target is increased significantly (Supplementary Figure S4). In contrast, the binding of M. *Taq*I to its target DNA increased the width of only

the minor groove by 3 Å (34). Structurally, the additional residue required to plug the DNA is accommodated in a long loop (R1115 to I1141). F1133, discussed above, is also in this loop. In the prototypical M.*Taq*I, this loop is much shorter and lacks the two intercalating residues but has the conserved phenylalanine involved in T-interaction with the adenine.

Structural elements involved in target recognition

The structures of LlaGI and LlaBIII bound to their target sites provided a means to understand the structural basis of target recognition by Type ISP enzymes. LlaGI recognizes the 7 bp target sequence CTnAGYG, while LlaBIII recognizes the 6 bp target sequence TnGACC. The overall mode of target binding by the two enzymes is very similar, involving DNA bending and partial unwinding resulting in widening of both the major and minor groove at the target, and flipping of the adenine at +1 position into the MTase catalytic pocket to be methylated. Interestingly, target sequence specificity is achieved by two distinct means—(i) by substitution of a positionally conserved amino acid to ensure correct reading of a target base and (ii) through base-specific interactions by amino acids located at positionally distinct positions (Figure 6A and B). Nonetheless, the LlaGI and LlaBIII residues involved in target recognition are located in one of the following conserved structural elements (Figure 5B, C and 6B): Loop I (1017–1027; residue numbers based on LlaGI sequence), Loop II (1053–1062) and Loop III (1115–1138) in the MTase domain; and Loop IV (1223–1230), Loop V (1324–1331) and Loop VI (1365–1374) in the TRD (Figure 6B). Particular loops are involved in recognition of particular sequence positions. In addition, LlaGI utilizes a helix (H) spanning residues 1286–1294 in the TRD. The equivalent helix in LlaBIII does not play a recognition role. Figure 6A and B reveals the hydrogen-bonded interactions that the residues of LlaGI and LlaBIII use for readout of the different targets.

Evolutionary constraints at the target-binding region

The above comparison revealed that though not all the amino acids involved in target recognition of the disparate LlaGI and LlaBIII targets are positionally equivalent, all but one amino acid are located within one of the six structurally conserved loops. The conservation scores of the amino acids in the target-binding region derived from the MSA were comparatively lower (Figures 5B, C and 6C). At first glance, this indicated that the structural elements and amino acid positions used by LlaGI and LlaBIII for target recognition are not functionally conserved amongst other Type ISP enzymes. Alternatively, the lower conservation score could be a result of substitutions occurring at these positions to facilitate recognition of different target sequences. If latter were true, then the position, rather than the identity of the amino acid, should be under strict EC, and would co-evolve with one or more positions in the protein. Accordingly, there would also be a correlation between the amino acids at these positions and the target sequence.

Using the program EVcouplings (32), an evolutionary coupling score was calculated for all pairs of residues from 868 to 1428 (LlaGI numbering), which encompasses the target binding MTase domain, core and jaw. Based on the number of pairs a residue was associated with and the corresponding coupling scores, the program calculated strength of EC for that residue, as defined by Hopf *et al.* (2012) (41). It has been shown that the EC strength of a residue is linked to its functional importance, such as binding to ligands (32). Through this calculation, we sought to find if the residues of LlaGI and LlaBIII involved in target recognition were evolutionarily constrained. The top 10% of residue positions, i.e. 56 of 561, with a high EC strength is shown in Figure 6C (for the complete list see Supplementary Data 2). Out of the 17 residues of LlaGI and/or LlaBIII that participate in direct base-specific interactions, 11 were in the top 5% and 3 more within the top 10%. Among the remaining three positions, 1023 (LlaGI numbering) was within top 15%, while 1024 and 1055 were below the top 65% (Supplementary Data 2).

In the top 10%, 28 residues, though not involved in base-specific interactions, were still part of the structural elements involved in target recognition, while 11 of them formed a single β-hairpin loop (1347–1358) in the core subdomain. The two strands of the hairpin are connected by a Ω loop, and make extensive interactions with other structural elements including Loop VI, in addition to directly or indirectly interacting with the phosphate backbone of the DNA. This loop appears structurally conserved in M.*Taq*I, Type I HsdS subunits and TRD of BpuSI. Based on the conservation and its location, we speculate that the loop may impart structural stability to the core.

Correlation between amino acids and target sequence

Having found that the structural elements involved in target recognition are under strong EC, we proceeded to find if there is a correlation between the target sequence and the amino acid substitutions at positions involved in its recognition. In addition to LlaBIII and LlaGI, the targets of nine other Type ISP enzymes are known (Figure 7A) (20). (In addition, Mtu10134II, MboBCG21III and MtuHN878II have the same amino acid sequence as Mtu18II and recognize the same target sequence.) We compared the amino acid and the target sequences of these enzymes along with LlaGI and LlaBIII to gain insights into this question and investigate if there is a consensus code for target recognition (Figure 7B). The sequence identity amongst the MTase-TRD unit of the 11 enzymes ranged from ∼87% (HpyAXVIII versus HpyUM037X) to ∼33% (HpyUM037X versus LlaBIII). The summary of the analyses is given below and is elaborated in the Supplementary Text.

Position −1: LlaGI recognizes C:G (CTNGAYG). Cytosine and guanine interact with Q1118 and K1131 of Loop III, respectively, via the major groove (Figure 6A). The sequence analysis revealed a very strong correlation between the occurrence of lysine at 1131 and C:G at −1 (Figure 7A). In enzymes that do not recognize −1, including LlaBIII, the glutamine is substituted by a small- or medium-sized hydrophobic residue and the lysine by asparagine (Figure 7A). We predict that a G:C at this position could be read by arginine at 1125 (see Supplementary Text for details).

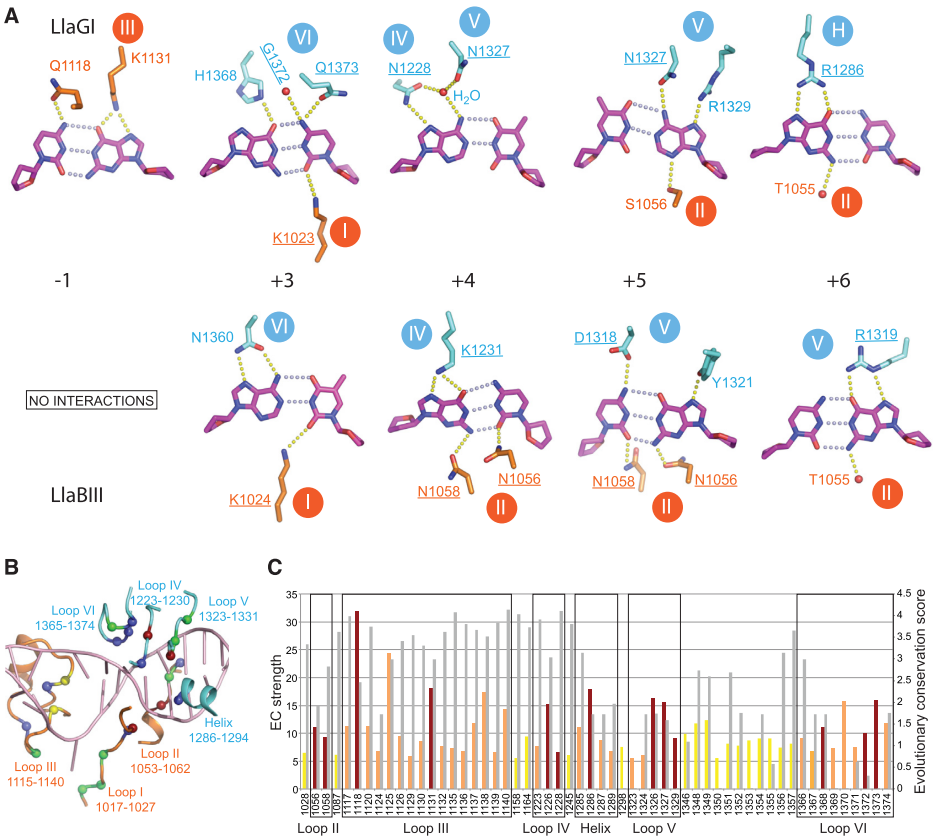


Figure 6. Target recognition. (**A**) Base-specific interactions made by LlaGI (top) and LlaBIII (bottom). Hydrogen bond acceptors and donors within 3.5 Å are connected by yellow dotted lines. Positionally non-equivalent residues that read the equivalent base pairs of LlaGI and LlaBIII targets are underlined. Note that the interactions made by LlaGIΔN with the base pair at +1 are illustrated in Figure 5E. (**B**) The structural elements involved in target recognition (residue numbers based on LlaGI sequence). The spheres represent amino acid residues whose side chains are involved in target recognition in case of LlaGI and LlaBIII. Residues involved in target recognition of LlaGI are coloured blue, residues involved in recognition of LlaBIII target are coloured red and those involved in recognition of both the targets are coloured green. The two intercalating residues that fill the cavity formed upon flipping of the adenine are coloured yellow. (**C**) The top 10% of the residues of the MTase-TRD unit (LlaGI residues 868–1428) with the highest EC strengths. The EC strengths (coloured bars) and the conservation scores (grey bars) for each of these residues are plotted. The EC strengths and the conservation scores were computed using EVcouplings (32,40) and ConSurf (31). The red coloured bars are the EC strengths of residues that interact with a target base; the orange bars are of residues that do not interact with the target but are part of the structural elements involved in target recognition; the yellow bars are of residues in other regions of the MTase-TRD unit. The residue numbering is based on LlaGI sequence. Residues LlaGI-K1023, LlaBIII-K1024 and LlaGI/LlaBIII-T1055, which interact with target bases, are not shown because their score is less than the top 10% (see text).

Position +1**:** By convention, +1 represents the location of the adenine that is methylated by the MTase domain. Consequently, all the enzymes recognize a T:A at this position (for example, LlaGI-CTNGAYG and LlaBIII-TNAGCC). In the LlaGIΔN and LlaBIII structures, specificity of the T:A base pair is ensured by the interactions made by the flipped adenine with the active site residues of the MTase domain located primarily on Loop I, which catalyse the transfer of methyl group (Figure 5E). It is possible that enzymes which modify cytosine at this position may be discovered, or even engineered in future.

Position +2**:** Among the 11 Type ISP enzymes whose targets are known, only three recognize this position in a sequence-specific manner. In both LlaGI and LlaBIII structures, the +2 base pair lies at the mouth of the MTase-TRD clamp. The bases do not make any sequence-specific contacts with the protein, thus making this position sequence non-specific. From the sequence alignment, it was not obvious how the other three enzymes read this position.

Position +3**:** LlaGI recognizes G:C (CTNGAYG) while LlaBIII recognizes A:T (TNAGCC) at this position. The top strand guanine of the LlaGI target and the adenine at the corresponding position of LlaBIII target are recognized by interactions made by the positionally equivalent residues, H1368 and N1360, respectively. The bottom-strand cytosine of the LlaGI target is recognized by the TRD residues G1372 and Q1373, and the MTase residue K1023. However, the corresponding bottom-strand thymine of the LlaBIII target interacts only with the LlaBIII MTase residue K1024 (Loop I).

Position +4**:** LlaGI recognizes A:T (CTNGAYG) while LlaBIII recognizes G:C (TNAGCC) at this position. The A:T base pair at the LlaGI target is recognized by the interaction between N1228 with the adenine. A water molecule assigned based on a weak $2F_O$-F_C but a strong F_O-F_C electron density bridged interactions between the base and N1228 and N1327 (Figure 6A). In LlaBIII, the guanine at the top strand interacts with K1231 and N1058, and the complementary cytosine interacts with the MTase residue

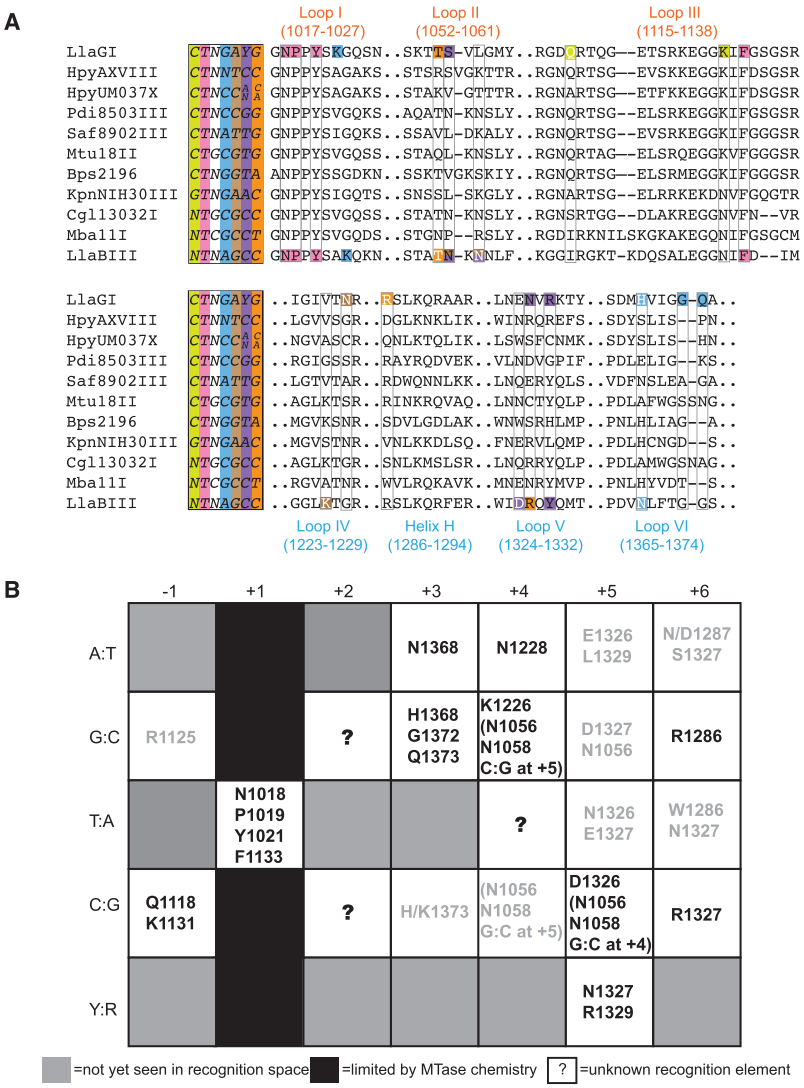


Figure 7. Consensus amino acid code for target recognition. (**A**) An MSA of the amino acid sequences of the seven structural elements involved in target recognition. The sequences are of those Type ISP enzymes whose targets are known. On the left of the alignment are the name of the enzyme and the target sequence. Each position of the target is coloured distinctly. The residues involved in direct interactions with the target bases of LlaGI and LlaBIII are coloured according to the position of the base they interact with. Residues that interact with the top strand are in white font, while those that interact with the bottom strand are in black. The residue numbering is based on LlaGI sequence. (**B**) A tabulation of the amino acids of the Type ISP MTase-TRD involved in recognition of the different base pairs at six of the seven target positions. Amino acids that interact with two neighbouring bases of a base pair step are in parenthesis. The neighbouring base pair, which is part of the step, and its position is also mentioned. Amino acids predicted by us to interact with target bases are in grey font.

N1056 via the minor groove. An A:T base pair at +4 correlated with asparagine at 1228, while G:C correlated strongly with lysine at 1226 (Figure 7A). We noticed that a GC:GC or CG:CG base pair step at positions +4 and +5 correlated with asparagine at 1056 and 1058.

Position +5: LlaGI recognizes T:A or C:G (CTNGAYG). In the crystal structure, the DNA bound to LlaGI has T:A at this position. The bottom strand adenine interacts with the TRD residues N1327 and R1329, and the MTase residue S1056. At this position, LlaBIII recognizes only C:G (TNAGCC). Specificity for cytosine is achieved through contact with the LlaBIII residue D1318. LlaBIII has asparagine at 1056 and 1058, which interact with the guanine and cytosine via the minor groove, respectively, and LlaBIII-Y1321 interacting with guanine via the major groove. Based on the sequence alignment of 11 Type ISP enzymes whose targets are known, we predict that asparagine at 1326 or glutamate at 1327 could contribute to recognition of T:A; glutamate at 1326 could contribute to recognition of A:T; aspartate at 1327 could read G:C at +5 (see Supplementary Text for details).

Position +6: LlaGI recognizes a G:C (CTNGAYG) while LlaBIII recognizes a C:G (TNAGCC). The top strand guanine of LlaGI target is recognized through a bidentate interaction with R1286 located on helix H and the interaction by the backbone carbonyl of the MTase residue T1055 (Figure 6A). At the corresponding position of LlaBIII target, the bottom strand guanine is recognized by bidentate hy-

drogen bonds with LlaBIII-R1319 in Loop V (Figure 6A). We found a strong correlation between G:C at +6 and arginine at 1286, and C:G and arginine at 1327 (Figure 7A). Based on our analyses we predict that recognition of A:T at +6 could involve aspartate at 1287 and serine at 1327; T:A at +6 by tryptophan at 1286 and asparagine at 1327 (see Supplementary Text for details).

DISCUSSION

The structure determination and analysis of LlaGI along with the structure of LlaBIII and amino acid sequences of other Type ISP enzymes, described above, provided us with valuable insights into the recognition of target sequence. It has been a long-standing endeavour to rationally design restriction enzymes with new specificities. Single-polypeptide restriction enzymes with fused nuclease and methyltransferase are good candidates for engineering new specificities, as change in specificity alters simultaneously the targets of both the nuclease and methyltransferase (9). Perplexingly, irrespective of the enzyme system, strategies such as rational replacement of amino acids contacting the bases have only been partially successful and required multiple attempts to find active enzymes (8–10).

The comparison of target recognition by LlaGI and LlaBIII provided an alternative to understanding this issue, as we could compare two very closely related RM enzymes that recognize different target sequences. Upon comparison, we found that most of the structural elements employed by the two enzymes for target recognition were identical. However, changes in target sequence did not always involve appropriate substitution of amino acids at the same positions. Corresponding amino acid changes also occurred at positionally distinct locations, with possible compensatory mutations in regions beyond to maintain the integrity of the structural elements. The non-equivalent positions were either on the same or one of the other target-binding structural elements. In addition, indirect interactions, such as the water bridge between LlaGI-N1228 and adenine at +4, appeared to contribute to specificity. These results emphasize the complexities of target recognition and that it may not suffice to change the identity of amino acids contacting the target bases to generate new specificity. Additional substitution of amino acids at new positions may be required to read the new set of bases.

Analyses of the MSA of a large set of Type ISP enzyme sequences revealed that the amino acid positions that are involved in target recognition are not strongly conserved. The lower conservation possibly reflects the change in amino acid identity complementary to the target sequence. Among the target binding structural elements, those belonging to the MTase domain were found to be strikingly more conserved than those from TRD (Figures 5B, C and 6C). The MTase residues play crucial role in specificity of −1 via the major grove and the flipped adenine of +1. The contacts made by the relatively less conserved TRD residues are the primary sequence readers of the base pairs downstream from +3 onwards, while interactions by the MTase domain via the minor groove provide additional stability. The differential conservation score suggests positions −1 and +1 to be least plastic. This observation highlights the constraints in engineering the DNA binding elements that are part of or in the near vicinity of catalytic regions, in this case the MTase active site. Despite relatively lower conservation, we found that the residues of the structural elements recognizing the target, including those that interact with the bases, were under strong EC. Our analysis using the program EVcouplings identified the target recognition region as the primary functional 'hotspot' in the MTase-TRD unit, suggesting this to be a powerful tool in identifying specificity-determining regions in DNA binding proteins.

Based on our analysis, we have arrived at a consensus code for target recognition by Type ISP enzymes (Figure 7B). This, we believe, will help in engineering enzymes with new specificity. We recognize that the engineering of new specificity will be limited by the catalytic requirement of a T:A base pair at +1. It will be interesting to find if the binding pocket of the adenine methyltransferase can be engineered to accommodate and recognize other bases. The translocation-active nuclease deletion mutant of LlaGI reported here highlights the modular nature of these enzymes. The modularity of the Type ISP enzymes is an additional advantage in engineering sequence-specific methyltransferases, DNA translocases or nickases. These would be useful biotechnological reagents where N6-adenine methylation is to be carried out; sequence-dependent remodelling of nucleoprotein complexes is required; or sequence-specific nicking of dsDNA is to be performed. The conserved architecture of the MTase-TRD unit of the γ-class of N6-adenine methyltransferase, Type I, Type IIG and L enzymes suggests that the insights on target recognition by Type ISP enzymes may find use in engineering specificity in the above enzyme types too.

ACCESSION NUMBER

The atomic coordinates and structure factors have been deposited with accession code 5FFJ.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank European Synchrotron Radiation Facility (ESRF), Grenoble, France, and Diamond Light Source (DLS), Oxfordshire, UK, for access to their beamlines. We also thank DBT for funding the use of BM14 beamline at ESRF. We would like to thank the anonymous reviewers for their constructive comments on the manuscript.

FUNDING

Wellcome Trust-DBT India Alliance [500048-Z-09-Z to K.S.]; Wellcome Trust [084086 to M.D.S., K.v.A.]; CSIR India (to N.N.). Funding for open access charge: Wellcome Trust-DBT India Alliance.

Conflict of interest statement. None declared.

REFERENCES

1. Luscombe,N.M., Laskowski,R.A. and Thornton,J.M. (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.*, **29**, 2860–2874.
2. Chandrasegaran,S. and Carroll,D. (2015) Origins of programmable nucleases for genome engineering. *J. Mol. Biol.*, doi:10.1016/j.jmb.2015.10.014.
3. Filipovska,A., Razif,M.F.M., Nygard,K.K.A. and Rackham,O. (2011) A universal code for RNA recognition by PUF proteins. *Nat. Chem. Biol.*, **7**, 425–427.
4. Coquille,S., Filipovska,A., Chia,T., Rajappa,L., Lingford,J.P., Razif,M.F.M., Thore,S. and Rackham,O. (2014) An artificial PPR scaffold for programmable RNA recognition. *Nat. Commun.*, **5**, doi:10.1038/ncomms6729.
5. Takeuchi,R., Choi,M. and Stoddard,B.L. (2015) Engineering of customized meganucleases via in vitro compartmentalization and in celluo optimization. *Methods Mol. Biol.*, **1239**, 105–132.
6. Dorner,L.F., Bitinaite,J., Whitaker,R.D. and Schildkraut,I. (1999) Genetic analysis of the base-specific contacts of BamHI restriction endonuclease1. *J. Mol. Biol.*, **285**, 1515–1523.
7. Lukacs,C.M., Kucera,R., Schildkraut,I. and Aggarwal,A.K. (2000) Understanding the immutability of restriction enzymes: crystal structure of BglII and its DNA substrate at 1.5 A resolution. *Nat. Struct. Mol. Biol.*, **7**, 134–140.
8. Townson,S.A., Samuelson,J.C., Xu,S.Y. and Aggarwal,A.K. (2005) Implications for switching restriction enzyme specificities from the structure of BstYI bound to a BglII DNA sequence. *Structure*, **13**, 791–801.
9. Morgan,R.D. and Luyten,Y.A. (2009) Rational engineering of type II restriction endonuclease DNA binding and cleavage specificity. *Nucleic Acids Res.*, **37**, 5222–5233.
10. Morgan,R.D., Luyten,Y.A., Johnson,S. and Wilson,G.G. (2015) Comprehensive Analysis of DNA Sequence Specificity. In: *7th New England Biolabs Meeting on DNA Restriction and Modification, August 24-29, 2015*. Univ. of Gdansk, Poland, p. 42.
11. Chand,M.K., Nirwan,N., Fiona,D., Kulkarni,M., Pernstich,C., Szczelkun,M. and Saikrishnan,K. (2015) Translocation-coupled DNA cleavage by the Type ISP restriction-modification enzymes. *Nat. Chem. Biol.*, **11**, 870–877.
12. Madsen,A. and Josephsen,J. (2001) The LlaGI restriction and modification system of Lactococcus lactis W10 consists of only one single polypeptide. *FEMS Microbiol. Lett.*, **200**, 91–96.
13. Smith,R.M., Diffin,F.M., Savery,N.J., Josephsen,J. and Szczelkun,M.D. (2009) DNA cleavage and methylation specificity of the single polypeptide restriction – modification enzyme LlaGI. *Nucleic Acids Res.*, **37**, 7206–7218.
14. Smith,R.M., Josephsen,J. and Szczelkun,M.D. (2009) An Mrr-family nuclease motif in the single polypeptide restriction – modification enzyme LlaGI. *Nucleic Acids Res.*, **37**, 7231–7238.
15. Smith,R.M., Josephsen,J. and Szczelkun,M.D. (2009) The single polypeptide restriction – modification enzyme LlaGI is a self-contained molecular motor that translocates DNA loops. *Nucleic Acids Res.*, **37**, 7219–7230.
16. Sisakova,E., van Aelst,K., Diffin,F.M. and Szczelkun,M.D. (2013) The Type ISP restriction-modification enzymes LlaBIII and LlaGI use a translocation-collision mechanism to cleave non-specific DNA distant from their recognition sites. *Nucleic Acids Res.*, **41**, 1071–1080.
17. van Aelst,K., Sisakova,E. and Szczelkun,M.D. (2013) DNA cleavage by Type ISP Restriction – Modification enzymes is initially targeted to the 3′ -5′ strand. *Nucleic Acids Res.*, **41**, 1081–1090.
18. van Aelst,K., Saikrishnan,K. and Szczelkun,M.D. (2015) Mapping DNA cleavage by the Type ISP restriction-modification enzymes following long-range communication between DNA sites in different orientations. *Nucleic Acids Res.*, **43**, 10430–10443.
19. Rhodes,D., Schwabe,J.W., Chapman,L. and Fairall,L. (1996) Towards an understanding of protein-DNA recognition. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **351**, 501–509.
20. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2015) REBASE–a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, **43**, D298–D299.
21. Leslie,A. and Powell,H. (2007) Processing diffraction data with MOSFLM. In: Read,RJ and Sussman,JL (eds). *Evolving Methods for Macromolecular Crystallography*. Springer press, Dordrecht, pp. 41–51.
22. Kabsch,W. (2010) XDS. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **D66**, 125–132.
23. Evans,P.R. and Murshudov,G.N. (2013) How good are my data and what is the resolution? *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **D69**, 1204–1214.
24. Read,R.J. and McCoy,A.J. (2011) Using SAD data in Phaser. *Acta Crystallogr. Sect. D. Biol. Crystallogr.*, **67**, 338–344.
25. Murshudov,G.N., Skubák,P., Lebedev,A.A., Pannu,N.S., Steiner,R.A., Nicholls,R.A., Winn,M.D., Long,F. and Vagin,A.A. (2011) *REFMAC* 5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **67**, 355–367.
26. Adams,P.D., Afonine,P.V., Bunkóczi,G., Chen,V.B., Davis,I.W., Echols,N., Headd,J.J., Hung,L.-W., Kapral,G.J., Grosse-Kunstleve,R.W. *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **66**, 213–221.
27. Emsley,P., Lohkamp,B., Scott,W.G. and Cowtan,K. (2010) Features and development of Coot. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **66**, 486–501.
28. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
29. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Soding,J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, doi:10.1038/msb.2011.75.
30. Waterhouse,A.M., Procter,J.B., Martin,D.M.A., Clamp,M. and Barton,G.J. (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
31. Ashkenazy,H., Erez,E., Martz,E., Pupko,T. and Ben-Tal,N. (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.*, **38**, W529–W533.
32. Marks,D.S., Colwell,L.J., Sheridan,R., Hopf,T.A., Pagnani,A., Zecchina,R. and Sander,C. (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, **6**, e28766.
33. Buchan,D.W.A., Minneci,F., Nugent,T.C.O., Bryson,K. and Jones,D.T. (2013) Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.*, **41**, 349–357.
34. Goedecke,K., Pignot,M., Goody,R.S., Scheidig,a.J. and Weinhold,E. (2001) Structure of the N6-adenine DNA methyltransferase M.TaqI in complex with DNA and a cofactor analog. *Nat. Struct. Biol.*, **8**, 121–125.
35. Kim,J.-S., DeGiovanni,A., Jancarik,J., Adams,P.D., Yokota,H., Kim,R. and Kim,S.-H. (2005) Crystal structure of DNA sequence specificity subunit of a type I restriction-modification enzyme and its functional implications. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 3248–3253.
36. Shen,B.W., Xu,D., Chan,S., Zheng,Y., Zhu,Z., Xu,S. and Stoddard,B.L. (2011) Characterization and crystal structure of the type IIG restriction endonuclease RM. BpuSI. *Nucleic Acids Res.*, **39**, 8223–8236.
37. Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.
38. Viadiu,H. and Aggarwal,A.K. (2000) Structure of BamHI bound to nonspecific DNA: a model for DNA sliding. *Mol. Cell*, **5**, 889–895.
39. Loenen,W.A.M., Dryden,D.T.F., Raleigh,E.A., Wilson,G.G. and Murray,N.E. (2014) Highlights of the DNA cutters: a short history of the restriction enzymes. *Nucleic Acids Res.*, **42**, 3–19.
40. Cheng,X. and Roberts,R.J. (2001) AdoMet-dependent methylation, DNA methyltransferases and base flipping. *Nucleic Acids Res.*, **29**, 3784–3795.
41. Hopf,T.A., Colwell,L.J., Sheridan,R., Rost,B., Sander,C. and Marks,D.S. (2012) 3D structures of membrane proteins from genomic sequencing. *Cell*, **149**, 1607–1621.