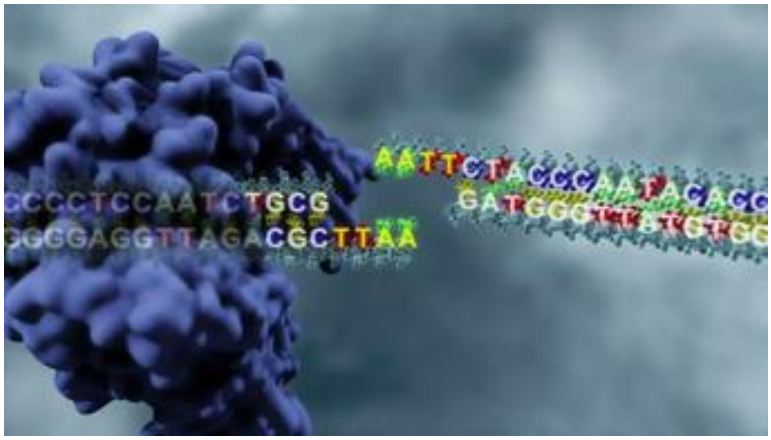
Univesity of Crete
BIOLOGY DEPARTMENT

FORTH
FOUNDATION FOR RESEARCH AND TECHNOLOGY - HELLAS
INSTITUTE OF MOLECULAR BIOLOGY AND BIOTECHNOLOGY

Master Thesis

Graduate Program of Protein Biotechnology

Identification and heterologous expression of CspAI Restriction and Modification system



Ioanna Makri

Scientific Supervisors: Prof.Michael Kokkinidis

Dr Dimitris Koutsioulis

MINOTECH BIOTECHNOLOGY

Heraklion 2018

Contents

Acknowledgements

First of all, I would like to express my deepest appreciation to my supervisor, Dr Dimitris Koutsioulis who trusted me and provided me the possibility to work on this project. Furthermore I would like to thank Professors Michale Kokkinidis and Vasilis Bouriotis as well as Associate Professor Dr Michael Aivaliotis for participating in my tripartite committee.

A special gratitude I give to Mrs Chrisoula, Mrs Chara and Mr Dimitris Dialektakis from Minotech who provided me all required equipment and necessary materials to complete this thesis. Special thanks go to Ioanna Kefala who supported me every day as well as to all members of laboratory, Mrs Pepi Karpeli, Dr Giannis Vlatakis, Sofia Markakiou and Panagiotis Kopsiaftis, for their advice and their willingness to help me every time I asked.

Last but not least, many thanks go to my family, who supports me in every possible way and stands by me for the 6 years of my studies.

Περίληψη

Οι περιοριστικές ενδονουκλεάσες είναι ένζυμα τα οποία αναγνωρίζουν και κόβουν μία συγκεκριμένη αλληλουχία DNA. Ο βιολογικός τους ρόλος είναι η προστασία του οργανισμού-ξενιστή από την εισβολή ενός ξένου DNA. Για να προστατευτεί το γενετικό υλικό του ξενιστή, εκφράζονται ένζυμα που τροποποιούν το DNA, οι μεθυλοτρανσφεράσες. Αυτά τα ένζυμα μεθυλιώνουν συγκεκριμένες αλληλουχίες DNA και τις προστατεύουν από την πέψη με τα αντίστοιχα περιοριστικά ένζυμα. Μία περιοριστική ενδονουκλέαση και η μεθυλοτρανσφεράση συνυπάρχουν και συνιστούν ένα RM (Restriction-Modification) σύστημα.

Οι περιοριστικές ενδονουκλεάσες που αναγνωρίζουν την ίδια αλληλουχία αλλά προέρχονται από διαφορετικά βακτηριακά στελέχη ονομάζονται ισοσχιζομερή. Πολλά ζεύγη ισοσχιζομερών διαφέρουν στην ευαισθησία τους στην μεθυλίωση. Για παράδειγμα, η CspAI και η AgeI αναγνωρίζουν την αλληλουχία 5'-A/CCGGT-3'. Όταν η υπογραμμισμένη κυτοσίνη είναι μεθυλιωμένη, η CspAI κόβει το DNA σε αντίθεση με την AgeI της οποίας η δράση αναστέλλεται. Ο τελικός μας στόχος είναι να συγκρίνουμε για πρώτη φορά τις κρυσταλλικές δομές των ισοσχιζομερών με διαφορετική ευαισθησία στη μεθυλίωση ώστε να προσδιορίσουμε τα κρίσιμα κατάλοιπα όσων αφορά την ευαισθησία στη μεθυλίωση. Αυτά τα δεδομένα θα βοηθήσουν τις μελλοντικές έρευνες στον λογικό σχεδιασμό των DNA-τροποποιητικών ενζύμων.

Η παρούσα εργασία εστιάζεται στον προσδιορισμό, την κλωνοποίηση και την έκφραση των συστατικών του CspAI RM συστήματος. Αρχικά, η αλληλούχιση ολόκληρου του γονιδιώματος του βακτηριακού στελέχους *Corynebacterium* αποκάλυψε τις αλληλουχίες DNA του περιοριστικού ενζύμου R.CspAI και της μεθυλοτρανσφεράσης M.CspAI. Και τα δύο γονίδια κλωνοποιήθηκαν και εκφράστηκαν σε επιλεγμένα στελέχη *E.coli* και η έκφραση της πρωτεΐνης R.CspAI ταυτοποιήθηκε με Western blot.

Οι στοιχίσεις πρωτεϊνικών αλληλουχιών του περιοριστικού ενζύμου CspAI με άλλες πρωτεΐνες έδειξαν ότι έχει τα ίδια κατάλοιπα στον καταλυτικό πυρήνα με την AgeI και ως εκ τούτου έχει το μοτίβο PD-(D/E)-XK. Περαιτέρω πειράματα θα αποκαλύψουν εάν ακολουθεί τον ίδιο μηχανισμό πέψης με την AgeI, ο οποίος είναι μοναδικός ανάμεσα στις Type IIP περιοριστικές ενδονουκλεάσες. Η αλληλουχία της μεθυλοτρανσφεράσης CspAI συγκρίθηκε με άλλες μεθυλοτρανσφεράσες, ψάχνοντας για συντηρημένα και χαρακτηριστικά μοτίβα και έδειξε υψηλή ομοιότητα με N4-N6 μεθυλοτρανσφεράσες.

Summary

Restriction endonucleases (REs) are enzymes which recognize and cleave a specific DNA sequence. Their biological function is to protect the host organism from invading DNA. In order to protect host's genomic DNA, DNA modification enzymes are expressed, methyltransferases (MT). These enzymes methylate specific DNA sequences and protect them from respective REs cleavage. RE and MT coexist and compose a RM system.

REs that recognize the same sequence but origin from different bacterial strains are called isoschizomers. Many isoschizomer pairs differ in their methyl-sensitivity. For example, CspAI and AgeI recognize sequence 5'-A/CCGGT-3'. When the underlined cytosine is methylated, CspAI cleave the DNA in contrast with AgeI that is blocked. Our ultimate goal is to compare for the first time the crystal structures of isoschizomers with different methyl sensitivity in order to identify the critical residues regarding methylation sensitivity. This data will help future studies on rational design of DNA modifying enzymes.

Current thesis focuses on identification, cloning and expression of CspAI RM system components. Primarily, whole genome sequencing of *Corynebacterium* revealed the DNA sequences of R.CspAI and M.CspAI. Both genes were cloned and expressed in selected *E.coli* strain and R.CspAI expression was verified by western blot.

Protein alignments of R.CspAI have demonstrated that it shares identical residues of the catalytic core with R.AgeI and hence, possesses the PD-(D/E)-XK motif. Further experiments will reveal if it has adapted the same cleavage mechanism with R.AgeI which is unique among Type IIP REs. M.CspAI sequence was compared against several MTs searching for conserved and characteristic motifs and showed higher similarities with N4-N6 MTs.

1. Introduction

1.1 Restriction enzymes

1.1.1 Biological role

Restriction endonucleases (or restriction enzymes, REs) are enzymes found in prokaryotes (bacteria and archaea) and in few viruses of unicellular algae. They act as an ‘’immune response’’, as their main function is to protect their host from invasion by foreign DNA, in particular bacteriophage. This defense mechanism includes the recognition of a specific DNA sequence and the cleavage of this sequence by the appropriate RE(A. Pingoud et al. 2005) (Figure 1).

In order to identify and cleave only the invading DNA, these organisms have developed a modification for protection of the genomic DNA. This modification is a methylation of either adenine (N6) or cytosine (C5 or N4) within the recognition site. These two processes, restriction and modification, form the RM systems (A. Pingoud et al. 2005).

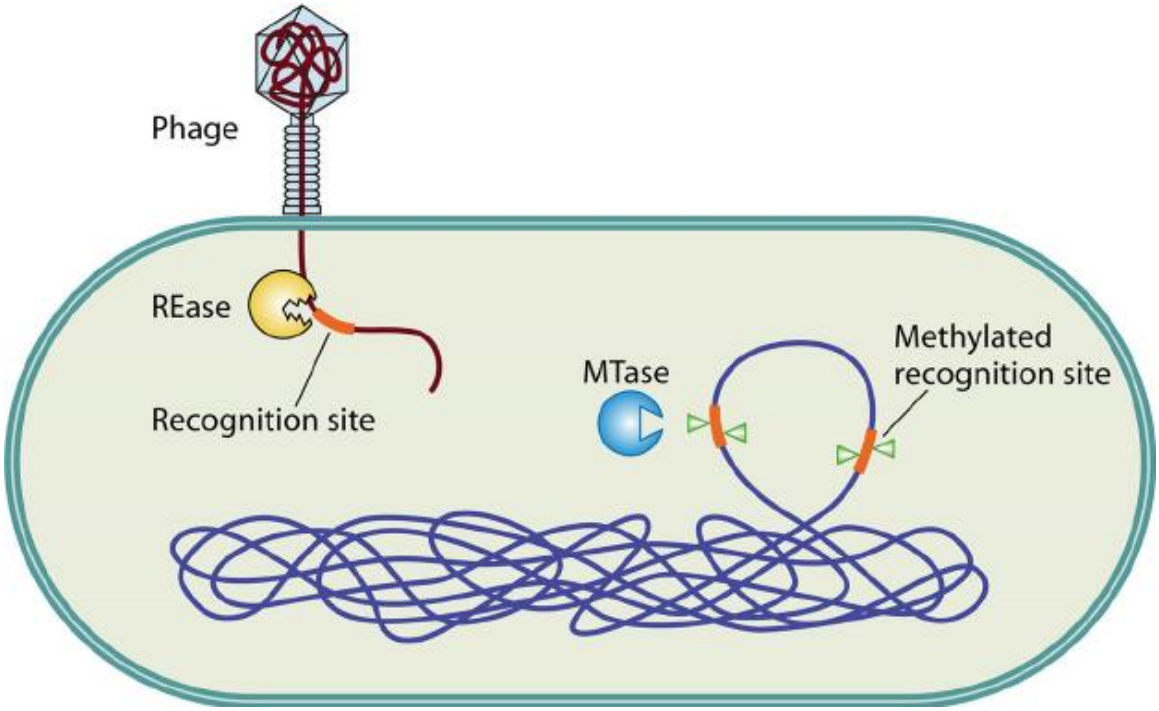


Figure 1: *The foreign bacteriophage’s DNA invades bacterial cell and the RE recognizes a specific sequence so as to cut it, whereas the respective methyltransferase has in advance modified the host DNA in order to protect it from cleavage. (Vasu and Nagaraja 2013)*

1.1.2 Classification

REs are classified in four main types, Type I, II, III and IV, according to subunits composition, recognition and cleavage sequence and requirement of cofactors (Table 1). Almost all types need divalent metal, usually Mg2+, for activity. In brief, Type I are oligomeric proteins, combined in a complex by 3 subunits, HsdR (restriction), HsdM (modification) and HsdS (specificity), which require hydrolysis of ATP for restriction and cut variably ,usually far from recognition site. Type II are separated or combined RE and methyltransferase and they cleave within or at fixed position close to target site. Type III form a complex with RE and methyltransferase (MT) and they demand ATP for restriction activity and sometimes SAM is needed. They cleave at specific position 25-27 nucleotides far from recognition site. Last but not least, Type IV are methylation-dependent REs, so they cleave at m6Adenine, m5Cytosine, hydro-m5Cytosine or other modified DNA but they cut at variable distance from recognition site (Loenen et al. 2014).

Type	Type I	Type II	Type III	Type IV
Features	Oligomeric REase and MTase complex Require ATP hydrolysis for restriction Cleave variably, often far from recognition site 'DEAD-box' translocating REase bipartite DNA recognition domain	Separate REase and MTase or combined REase~MTase fusion Cleave within or at fixed positions close to recognition site Many different subtypes	Combined REase+MTase complex ATP required for restriction Cleave at fixed position outside recognition site 'DEAD-box' REase	Methylation-dependent REase Cleave at variable distance from recognition site Cleave m6A, m5C, hm5C and/or other modified DNA Many different types
Example	e.g. EcoKI	e.g. EcoRI	e.g. EcoP1I	No 'typical' example
Genes	*hsdR, hsdM, hsdS*	e.g. *ecorIR, ecorIM*	e.g. *ecoP1IM, ecoP1IR*	e.g. *mcrA, mcrBC, mrr*
Subunits	~135, ~62 and ~52 kDa	~31 and ~38 kDa for EcoRI	~106 and ~75 kDa for EcoP1I	Unrelated proteins
Proteins	REase: 2R + 2M + S MTase: 2M+S (±2R)	Orthodox REase: 2R Orthodox MTase: M	REase: 1 or 2 R+2M MTase: 2M (±2R)	Varies
REBASE	104 enzymes, 47 genes cloned, 34 genes sequenced, 5140 putatives	3938 enzymes, 633 genes cloned, 597 sequenced, 9632 putatives	21 enzymes, 19 genes cloned & sequenced, 1889 putatives	18 enzymes & genes cloned, 15 sequenced, 4822 putatives

Type I and II are currently divided in 5 and 11 different subclasses, respectively. Few enzymes have been well-characterized, but based on the current avalanche of sequence information many putative genes belonging to all Types and subtypes are being identified and listed on the restriction enzyme website (http://rebase.neb.com). The modification-dependent Type IV enzymes are highly diverse and only a few have been characterized in any detail. In each case, an example is given of one of the best-characterized enzymes within the different Types I, II and III. Note that Type II enzymes range from simple (shown here for EcoRI) to more complex systems (see Table 2 for the diversity of Type II subtypes).
REBASE count is as of 16 September 2013 (http://rebase.neb.com/cgi-bin/statlist).

Table 1: Classification of REs according to their subunits composition, recognition and cleavage sequence and cofactor requirement (Loenen et al. 2014).

1.1.3 Type II REs: The molecular biology toolbox.

Type II REs are found in prokaryotes (bacteria and archaea) and in some viruses infecting eukaryotic algae, such as *Chlorella* (Alfred Pingoud, Wilson, and Wende 2014). The first characterized Type II RE was HindII from *Haemophilus influenza*, followed by EcoRI and EcoRII from *Escherichia coli* (Alfred Pingoud, Wilson, and Wende 2014)

Type II REs have been studied extensively because of their contribution to the development of recombinant DNA technology in molecular biology and medicine (Figure 2). This group is the largest one of identified REs with over 3,500 Type II enzymes been discovered and characterized, recognizing some 350 different DNA sequences (Loenen et al. 2014). Thousands more 'putative' Type II enzymes have been identified by analysis of sequenced bacterial and archaeal genomes, but remain uncharacterized.

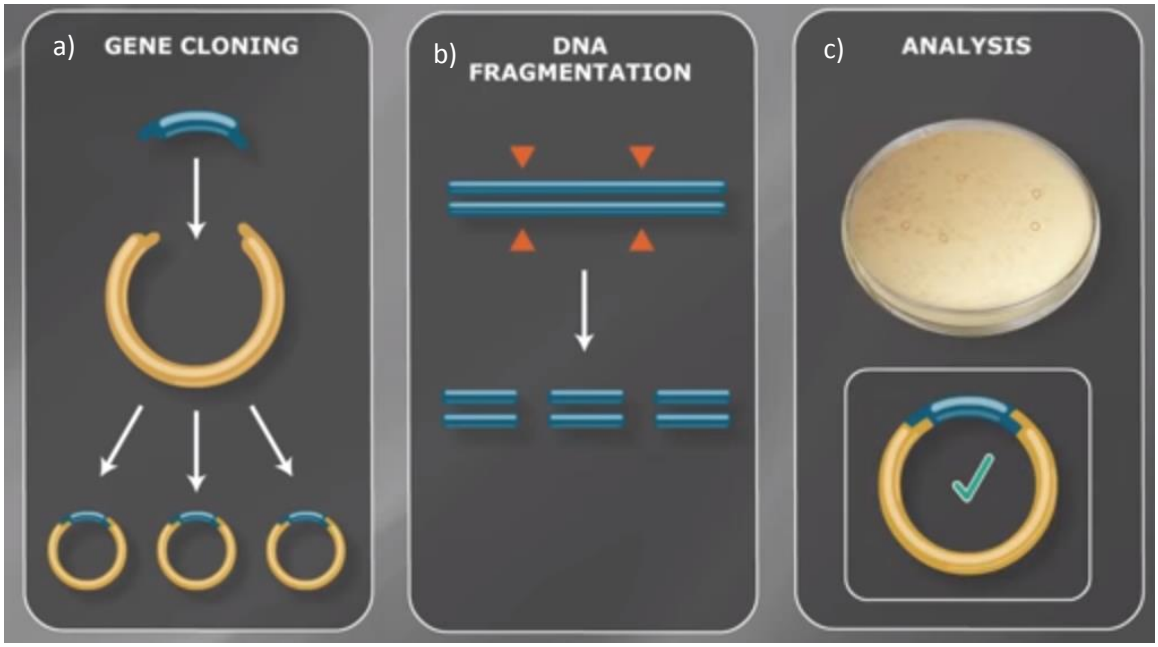


Figure 2: Type II REs are basic tools for most molecular biology applications such as a) gene cloning, b) cleavage of DNA and c) analysis of recombinant strains. (www.neb.com)

Their main function is to destroy the foreign DNA, providing a mechanism to protect the host from bacteriophages(Alfred Pingoud, Wilson, and Wende 2014). Type II REases cleave the double helix of DNA at a fixed position within or in close proximity to the recognition sequence, producing DNA fragments with a 3'-OH and a 5'-phosphate with sticky (5' or 3' overhangs) or blunt ends (Alfred Pingoud and Jeltsch 2001) (Figure 3). For the catalysis of the reaction, divalent cations ,usually Mg^{2+}, are necessary for almost all type II enzymes and they do not require ATP or GTP (Alfred Pingoud, Wilson, and Wende 2014).

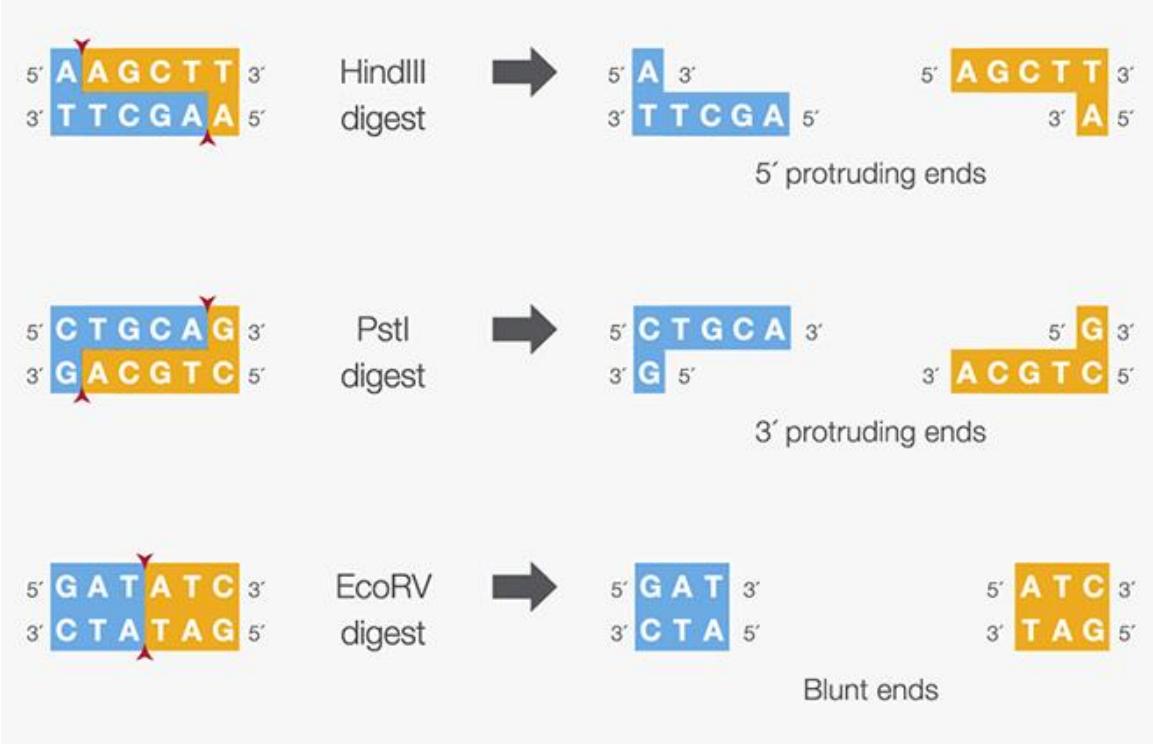


Figure 3: Illustration of the protruding or blunt ends produced by specific type II enzymes

To protect the host's DNA from cleavage *in vivo*, they coexist with modification enzymes, the MTs which occur as a monomer and act independently (Loenen et al. 2014).

Type II enzymes can be further categorized according to their enzymatic behavior in subclasses A, B, C, E, F, G, H, M, P, S and T. It should be considered not as an absolute classification, rather than that each group shows specific property. In fact, some Type II enzymes may belong to more than one subclass (Loenen et al. 2014) (Table 2), (Figure 4). For example, FokI belong to Type IIS subgroup but it could be classified also as IIE due to its requirement for binding in a second recognition sequence (Alfred Pingoud and Jeltsch 2001).

Subtype	Characteristic feature	Example[a]	Recognition sequence[b]
Orthodox	Palindromic recognition site, which is recognized by a homodimeric enzyme, cleavage occurs within or adjacent to the recognition site	***Eco*RI**	G↓ A A T T C C T T A A↑ G
		***Eco*RV**	G A T↓ A T C C T A↑ T A G
		***Bgl*I**	G C C N N N N↓ N G G C C G G N↑ N N N N C C G
Type IIS	Asymmetric recognition site with cleavage occuring at a defined distance	***Fok*I**	G G A T G N_9↓ N N N N C C T A C N_9 N N N N↑
Type IIE	Two sites required for cleavage, one serving as allosteric effector	***Nae*I**	G C G↓ C G C C G C↑ G C G
Type IIF	Two sites required for cleavage, both sites are cleaved in a concerted reaction by a homotetrameric enzyme	***Ngo*MIV**	G↓ C C G G C C G G C C↑ G
Type IIT	Different subunits with restriction and modification activity	*Bpu*10I	C C↓ T N A G C G G A N T↑ C G
		*Bsl*I	C C N N N N N↓ N N G G G G N N↑ N N N N N C C
Type IIG	One polypeptide chain with restriction and modification activity	*Eco*57I	C T G A A G N_{14} N N↓ G A C T T C N_{14}↑N N
Type IIB	Cleavage on both sides of the recognition site	*Bcg*I	N N↓ N_{10} C G A N_6 T G C N_{10} N N↓ ↑N N N_{10} G C T N_6 A C G N_{10}↑N N
		*Bpl*I	N N_4↓ N_8 G A G N_5 C T C N_8 N_4 N↓ ↑N N_4 N_8 C T C N_5 G A G N_8↑ N_4 N
Type IIM	Methylated recognition site	*Dpn*I	G ^{m}A↓ T C C T↑ $^{\mu}$A G

[a]Restriction endonucleases whose crystal structure is known are depicted in bold letters.
[b]The site of cleavage is indicated by ↓.

Table 2: Classification of type II REases in subclasses. The discrimination is not absolute as some type II enzymes belong to more than one subgroup (Alfred Pingoud and Jeltsch 2001).

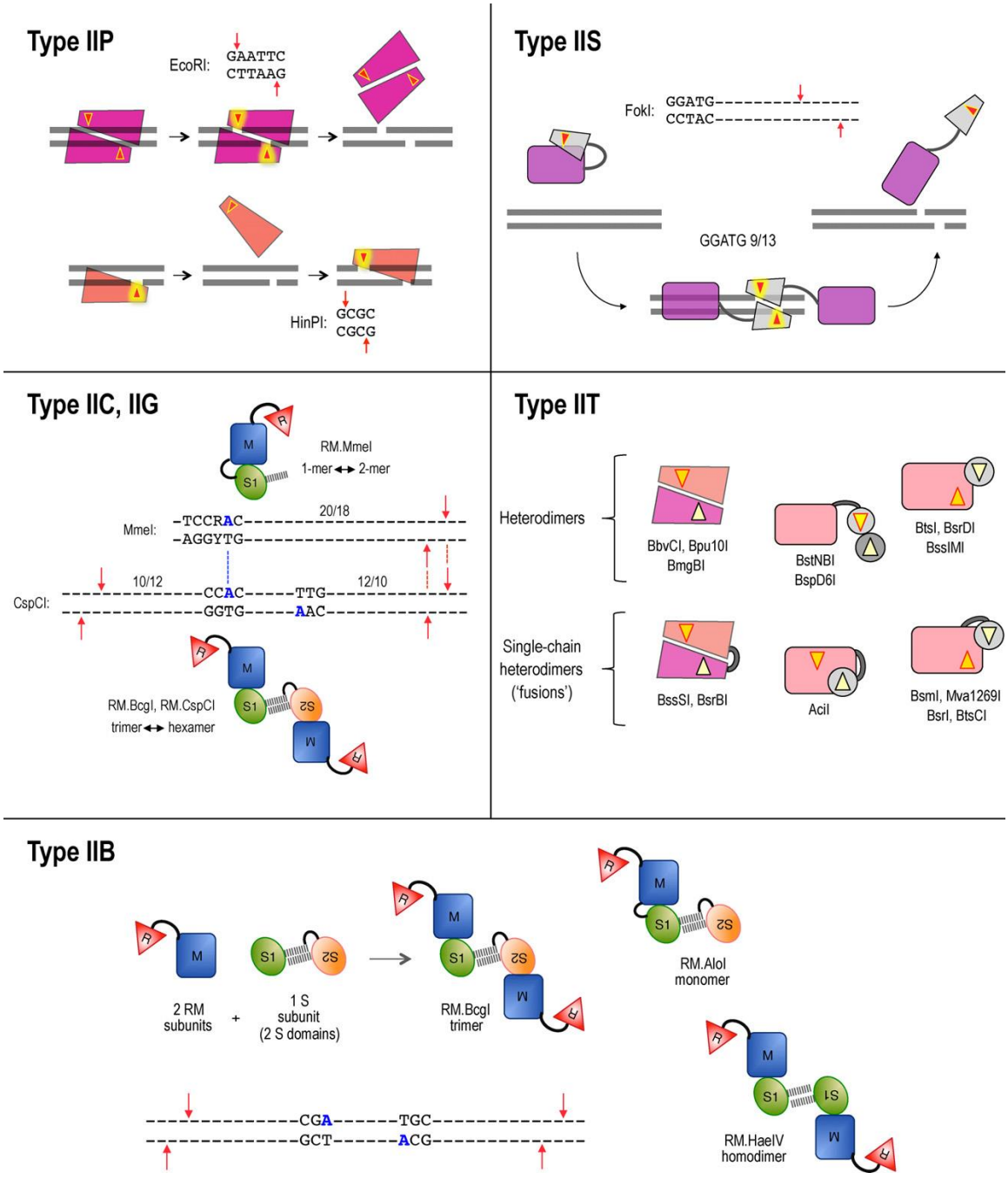


Figure 4: The subunit composition and cleavage mechanism of selected subtypes of Type II REases. Type IIP enzymes act mainly as homodimers (top), and cleave both DNA strands at once. Some act as dimers of dimers (homotetramers) instead, and do the same. Still others act as monomers (bottom) and cleave the DNA strands separately, one after the other. Bright triangles represent catalytic sites. Type IIS enzymes generally bind as monomers, but cleave as 'transient' homodimers. Type IIB enzymes cleave on both sides of their bipartite recognition sequences. Their subunit/domain stoichiometry and polypeptide chain continuity varies. Three examples of primary forms are shown: BcgI, AloI and HaeIV. These forms assemble in higher-order oligomers for cleavage. Type IIB enzymes display bilateral symmetry with respect to their methylation and cleavage positions. It is not clear whether they cleave to the left or to the right of the half-

sequence bound. Type IIG enzymes (e.g. BcgI) might cleave upstream (left) of their bound recognition half-site. All other Type IIG enzymes (e.g. MmeI) cleave downstream from the site, often with the same geometry. These proteins have very similar amino acid sequences, however, suggesting that somehow the reactions are the same. Type IIT enzymes cleave within or close to asymmetric sequences. Composition varies; they have two different catalytic sites: top-strand specific and bottom-strand specific. In some, both subunits/domains interact with the recognition sequence (left cartoons). In others, only the larger subunit/domain recognizes the DNA (Alfred Pingoud, Wilson, and Wende 2014).

The most familiar and unique subclass is Type IIP ("orthodox") which recognize palindromic sequences and cut symmetrically within the sequence or in a close proximity producing sticky or blunt ends (Alfred Pingoud, Wilson, and Wende 2014). The first Type IIP REs were EcoRI and EcoRV. They form mostly homodimers or homotetramers in order to be active but there are few enzymes which act as monomers (Figure 5). The structure depends on the length of the recognition sequence of the REs.

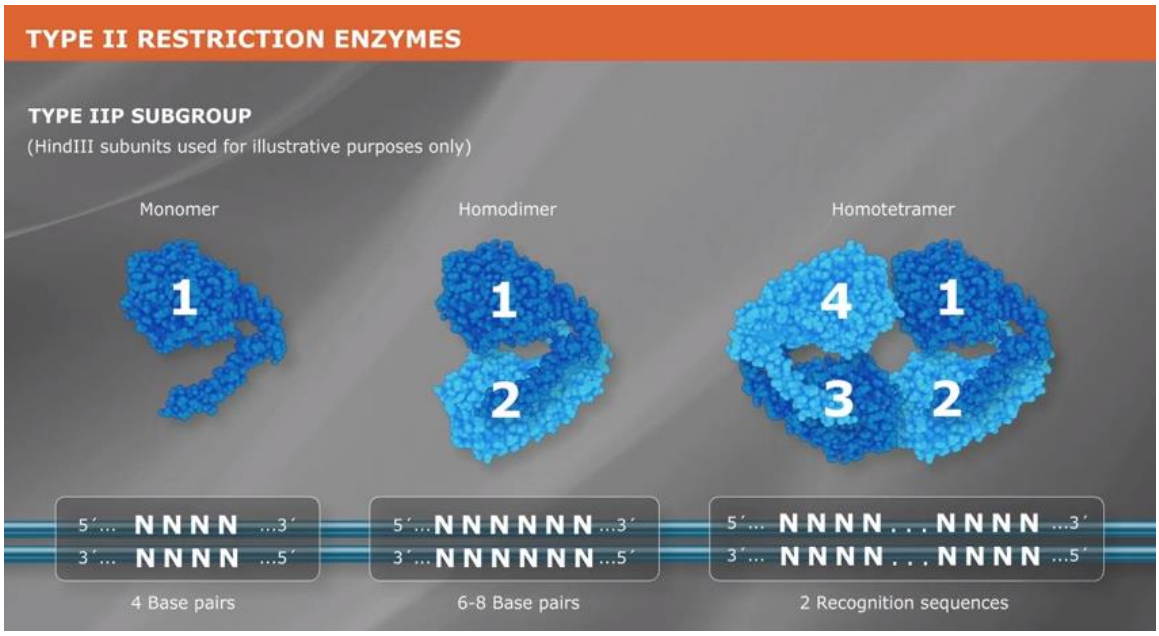


Figure 5: Type IIP enzymes act as monomers when the recognition site consists of 4 bp, as homodimer when the recognition site is composed of 6 to 8 bp and as homotetramer when they bind to and cleave 2 or more recognition sequence at once. (www.neb.com)

A motif PD-(D/E) XK was identified as common feature of type II restriction endonucleases. This motif appears also in other nucleases and it was identified as the catalytic core of Mg^{2+}-dependent nucleases (Loenen et al. 2014) . The motif consists of two or three acidic residues (Asp or Gly) and one Lysine residue which are combined to a bipartite catalytic motif. This motif is not absolute as some members of this family have developed different residues –variants- of the active site. All members of PD-(D/E)-XK family share a common structure consisting of a

mixed beta-sheet of four or five strands flanked on both sides by alpha-helices (Kosinski, Feder, and Bujnicki 2005) (Figure 6).

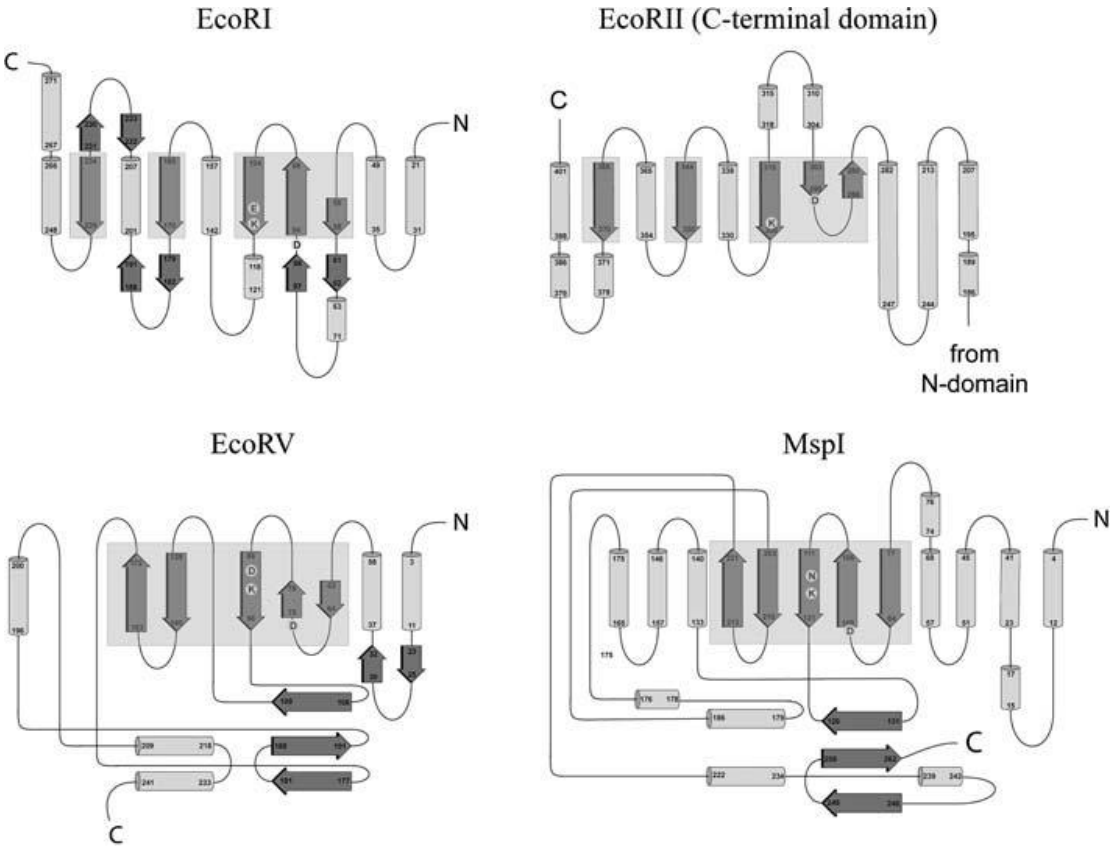


Figure 6: Secondary structure of four Type II REs, containing the conserved motif PD-(D/E)-XK. All four examples have a central five-stranded β-sheet (shaded grey) surrounding by α-helices (shown in light grey). The conserved motif PD-(D/E)-XK is located on the second and third strand of this β-sheet. β-strands are shown in dark grey (A. Pingoud et al. 2005).

1.1.4 Isoschizomers

Isoschizomers are pairs of REs which recognize and cleave the same DNA sequence but they are isolated from different strains of bacteria. There are also cases of isoschizomers which differ from each other according to their sensitivity in methylation of the target DNA sequence. One enzyme cleaves the appropriate sequence independently of the existence of a particular methylation of DNA, whereas the respective isoschizomer is blocked by the specific modification. A pair of isoschizomers, HpaII and MspI, is an example of this type of different methylation-sensitivity. Both enzymes recognize the sequence 5'-C/CGG-3' (cleavage is indicated by "/"). HpaII is the methyl-sensitive enzyme so it is unable to digest the target

sequence when the second cytosine is methylated, whereas MspI is not affected by the current methylation and catalyzes the cleavage (Kong, Application, and Data 2011).

Pairs of isoschizomers with difference in methylation sensitivity have become a valuable tool for epigenetics studies. DNA methylation of vertebrates is the most studied modification but it cannot be detected by PCR because it does not alter the DNA sequence. Thereby, REs with same recognition site but opposite methylation sensitivity made the discrimination between methylated and unmethylated locus easier for researchers. In fact, the isoschizomers HpaII and MspI have been used in order to discriminate whether a specific DNA sequence in mice cells is methylated or not (Khulan et al. 2006). There have been already developed methods based on the marks that isoschizomers produce (Figure 7)(BioLabs 2012).

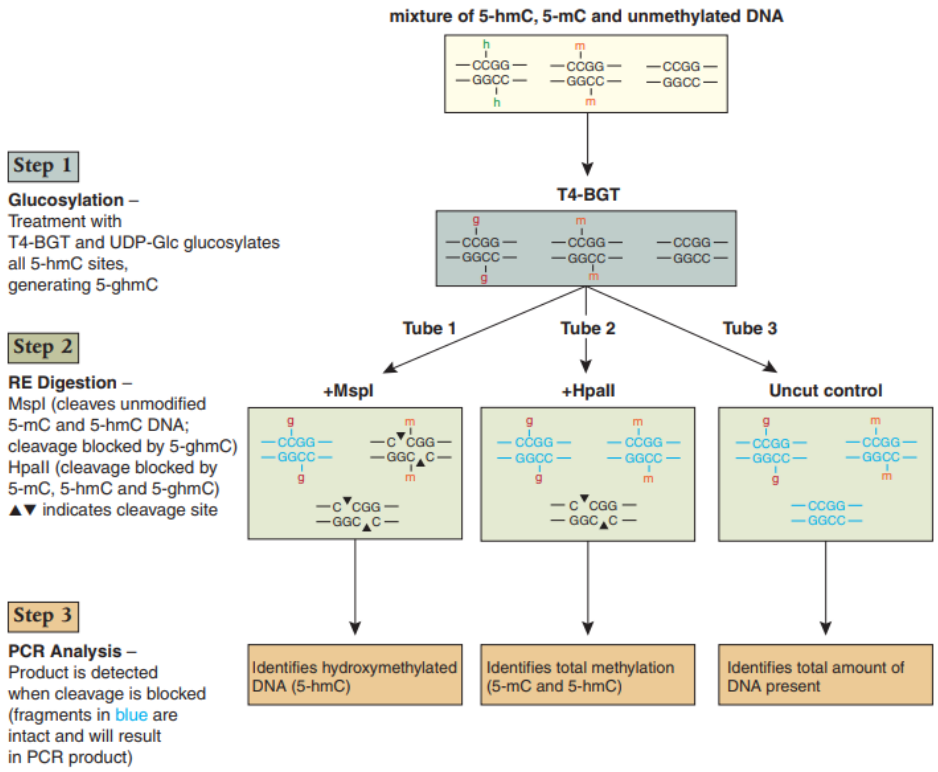


Figure 7: Experimental overview of EpiMark® 5-hmC and 5-mC Analysis Kit (New England Biolabs, Inc). The DNA of interest is treated with T4 β-Glucosyltransferase (T4-BGT) and UDP-Glucose (UDP-Glc). T4-BGT transfers glucose from UDP-Glc onto 5-hydroxymethylcytosine (generating glucosylated 5-hydroxymethylcytosine [5-ghmC]). MspI cuts DNA containing 5-hmC, but does not cut 5-ghmC containing sites; in contrast, HpaII is blocked by any of these modifications. Presence of 5-hmC and 5-mC can be determined by PCR analysis.

Apart from HpaII and MspI, there are many isoschizomer pairs which exhibit different sensitivity to DNA methylation modification. Examples of those isoschizomers are given in Table 3

Isoschizomer enzyme pairs

Recognition site	Methylation-insensitive	Methylation-sensitive
C**C**GG	MspI	HpaII
AC**C**GGT	CspAI	AgeI
CA**C**gN**C**GTC	DraIII	AdeI
CCCGGG	XmaI	SmaI
CCGCT**C**	BsrBI	MbiI
CY**C**GRG	BsoBI	AvaI
GANNNNTT**C**	XmnI	PdmI
GACGT**C**	ZraI	AatII
GAGCT**C**	SacI	EcI136II
GAGT**C**	MlyI	PleI
GAT**C**	CfuI	DpnI
GAWT**C**	TfiI	PfeI
GCNG**C**	BsoFI	SatI
G**C**gGC	BsoFI	Fnu4HI
GCNNNNNNNG**C**	MwoI	HpyF10VI
GCTAG**C**	BmtI	NheI
GG**C**C	HaeIII	NgoPII
GGCG**C**C	NarI	SfoI
3'-**C**CCTG	FaqI	BsmFI
GGNN**C**C	BspLI	NlaIV
GGTA**C**C	KpnI	Asp718I
GGW**C**C	AflI	Eco47I
GRCGY**C**	BsaHI	AhaII
GTA**C**	Csp6I	RsaI
GTCT**C**	Alw26I	BsmAI
3'-**C**AGAG	BsmAI	Alw26I
GTTTAAA**C**	MssI	PmeI
RC**C**GGY	BssAI	BsrFI
TC**C**GGA	AccIII	MroI
TT**C**GAA	SfuI	NspV

Table 3: Pairs of REs isoschizomers with different methylation sensitivity. In bold appears the cytosine which is methylated. All nucleotide sequence are shown in 5' to 3' orientation unless it is defined otherwise (Kong, Application, and Data 2011).

AgeI and CspAI is the isoschizomer pair to focus on as CspAI is the target enzyme of this study. They recognize and cut sequence 5'-A/CCGGT-3'. If the second cytosine is methylated, AgeI is

blocked whereas CspAI is active. The mechanism of cleavage of AgeI is unique and described below.

1.2. AgeI

AgeI is a RE isolated from *Agrobacterium gelatinovorum*, which recognizes the palindromic sequence 5'-A/CCGGT-3' and cleaves it at the position of A. It belongs to Type IIP subclass as well as to CCGG-family REases. Enzymes of this family contain a conserved motif R-(D/E)-R to recognize the tetranucleotide target site CCGG. They can act as dimers, tetramers or oligomers.

AgeI provide a unique cleavage mechanism for Type IIP but similar to that of Type IIS RE FokI. AgeI can be found in two forms: monomer and dimer. It exists as a monomer in its apo-form both in solution and in crystal structure but it dimerizes when it binds DNA. In the dimer form, the RE catalyzes the cleavage of DNA.

The apo-form is composed of 2 domains: a) N-terminal domain (1-82 residues) and b) C-terminal domain (83-278 residues). Once AgeI binds the duplex helix of DNA it forms a dimer and cleaves the specific sequence. The active site is constituted by the residues E97, D142, K168 and D178 (Figure 8). The composition of this active site is similar with BsaWI (which recognizes a related DNA sequence 5-W/CCGGW-3, W stands for A or T) and NgoMIV and other CCGG family enzymes. In particular, D178 of AgeI overlaps with D175 of BsaWI but does not match with the other members of CCGG family. AgeI structure is very similar to BsaWI structure (Figure 9). They share 24% identical and 41% similar amino acid sequence and their catalytic C-domains are similar.

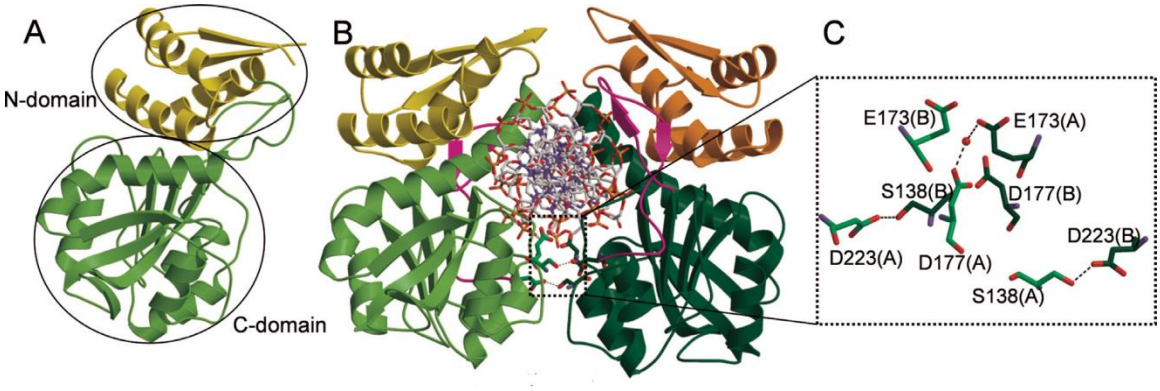


Figure 8: A) The apo-form of AgeI with the two domains, N- and C- domain. B) Crystal structure of AgeI in complex with the DNA. AgeI dimerizes when it binds with the DNA. N-domains are yellow and orange whereas C-domains are light and dark green. In magenta is depicted the element structure of AgeI (residues 197-224). C) An insight view of the residues that form the active site(Tamulaitiene et al. 2016).

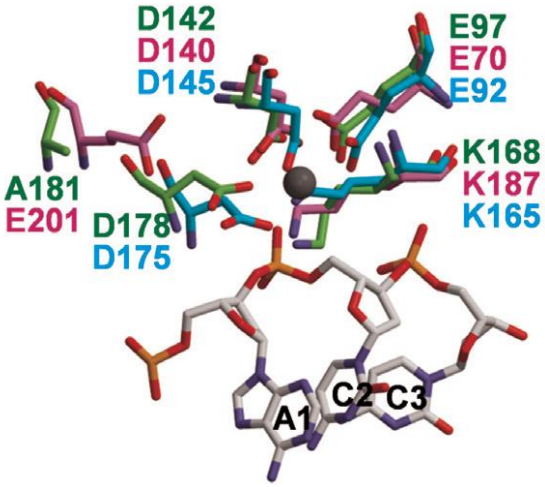


Figure 9: Active site residues of Age I (showed in green) overlaps with residues from catalytic core of BsaWI (showed in cyan) and with NgoMIV (showed in pink). D178 of AgeI overlaps only with D175 of BsaWI(Tamulaitiene et al. 2016).

AgeI recognize the CCGG tetranucleotide by its catalytic core. Although, it belongs to the CCGG family, AgeI possesses only the last two residues for the recognition of target site whereas the first arginine residue (R) is replaced by a lysine residue, K200. This arginine residue is responsible for the recognition of the first G from CCGG site. In crystal structure, AgeI contacts also with the minor groove. The Q86 residue of AgeI is responsible for the recognition of GG in the minor groove. This residue is also conserved in the CCGG family and in some case is replaced by N while its function is to contact the target site in the minor groove.

AgeI crystal structure shows a unique structure of the residues 197-224. This structure element, composed of K200, E214 and K224, recognizes the outer A: T bp (Figure 10).

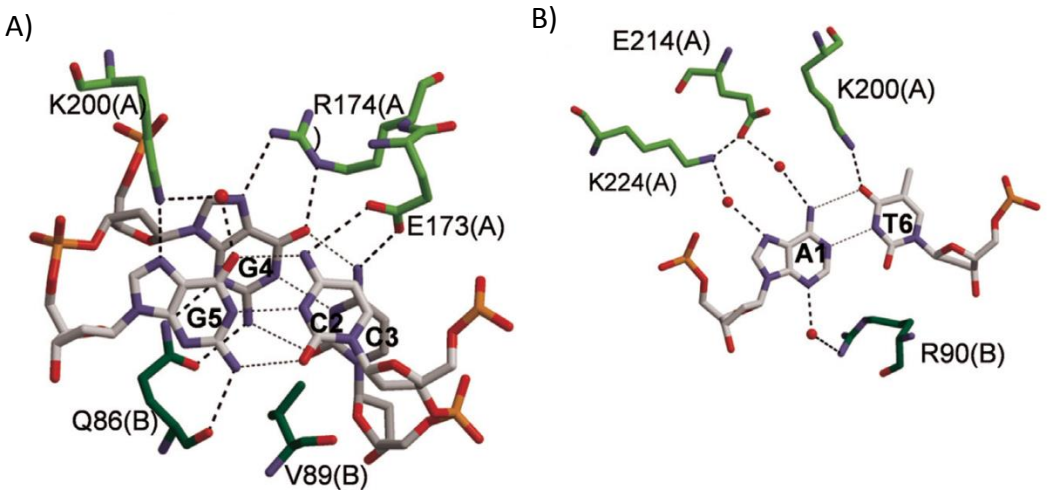


Figure 10: A) CC:GG recognition by AgeI, B) A:T recognition by the unique structure element(residues 197-224) of AgeI(Tamulaitiene et al. 2016).

1.2.1 Unique cleavage mechanism

AgeI forms a dimer when it binds to the DNA sequence. Each subunit interacts with the half site of the sequence 5'-ACCGGT-3' from the major groove and it contacts with another half site in the minor groove.

The dimerization is appropriate for catalysis of the reaction. In particular, each subunit cuts the phosphodiester bond on the opposite strand of the target sequence. Monomeric REs bind palindromic sites asymmetrically and cleave the strands sequential, whereas dimeric REs are dimers even in their apo-form. REs that exist as dimers in their apo form can create a tetramer or oligomer when they bind the double helix of DNA (Figure 11). DNA cleavage mechanism of Age I is more similar to that of Type IIS REs, FokI, which is a monomer in solution but after binding to the target site through the binding domain, FokI nuclease domains form a dimer and cleave both DNA strands at one of the target sites (Figure 11). Thereby, the AgeI cleavage-mechanism is unique among the REs.

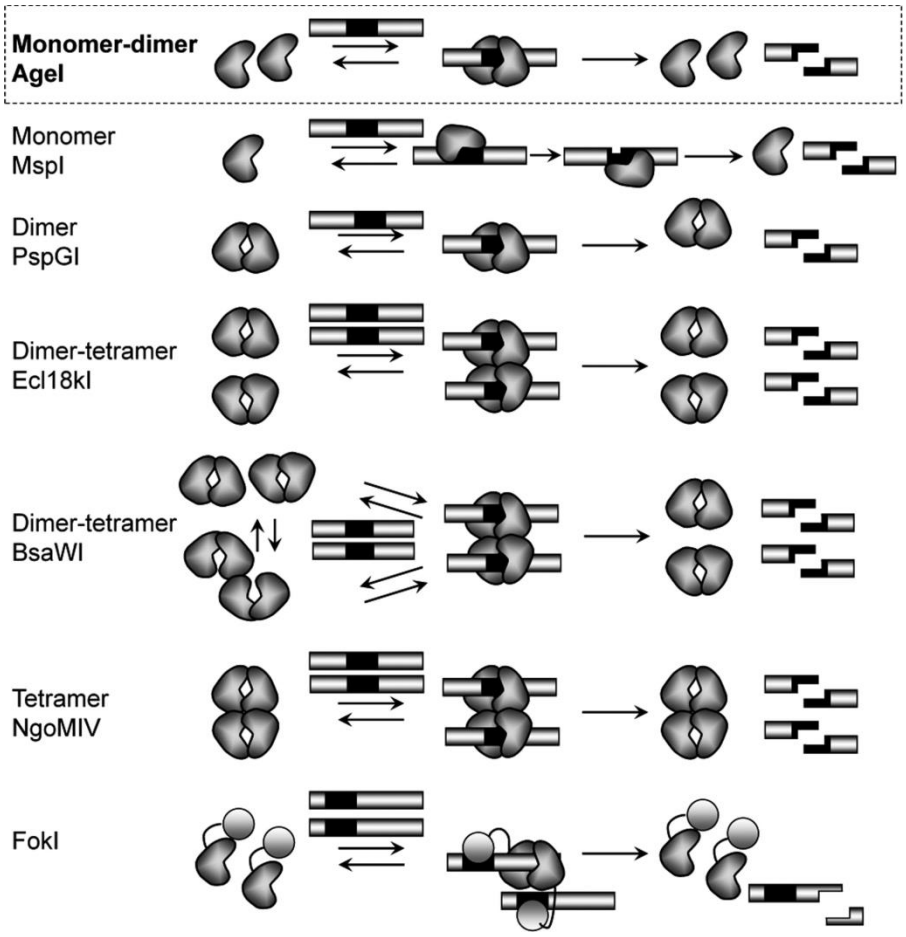


Figure 11: Oligomeric forms and DNA cleavage mechanisms of Type II REs interacting with (pseudo)palindromic sequences and Type IIS enzyme FokI. Black rectangles represent recognition sequence.

1.3 DNA methyltransferases

Modification of DNA by methylation occurs not only in prokaryotes, but in eukaryotes as well. In higher eukaryotes, cytosine methylation appears to participate in regulation of gene expression, embryonic development, genomic imprinting and X-chromosome inactivation. In prokaryotes, cytosine and adenine methylation are involved in restriction-modification systems ,whereas adenine methylation participate also in regulation of the initiation of DNA replication and targeting the correction of errors in DNA replication (Kumar et al. 1994).

Restriction-modification systems in prokaryotes involve the coexistence of MTs with respective REs and their main function is to protect host's DNA from cleavage.

The substrate of MTs is the DNA of the host cell ,which is modified at the recognition site and thus, protected against attack by the respective RE whereas the substrate of RE is the foreign/invading DNA, which is cleaved at specific sites (Alfred Pingoud and Jeltsch 2001).

MTs are classified in 3 classes: 2 classes modify exocyclic nitrogens , converting adenine to N6-methyladenine and cytosine to N4-methylcytosine and 1 class (endocyclic) modifies 5-carbon of pyrimidine converting it to 5-methylcytosine(Kumar et al. 1994). All known MTs utilize S-adenosyl-L-methionine (SAM) as the methyl-donor during methylation process. They transfer a methyl group (-CH3) from SAM to cysteine to produce AdoHcy(S-adenosyl-L homocysteine) and methylate cytosine or adenine residue of DNA (Bheemanaik, Reddy, and Rao 2006) (Figure 12). All three classes of MTs can be found in prokaryotes whereas only C5 MTs are found in eukaryotes (Bheemanaik, Reddy, and Rao 2006).

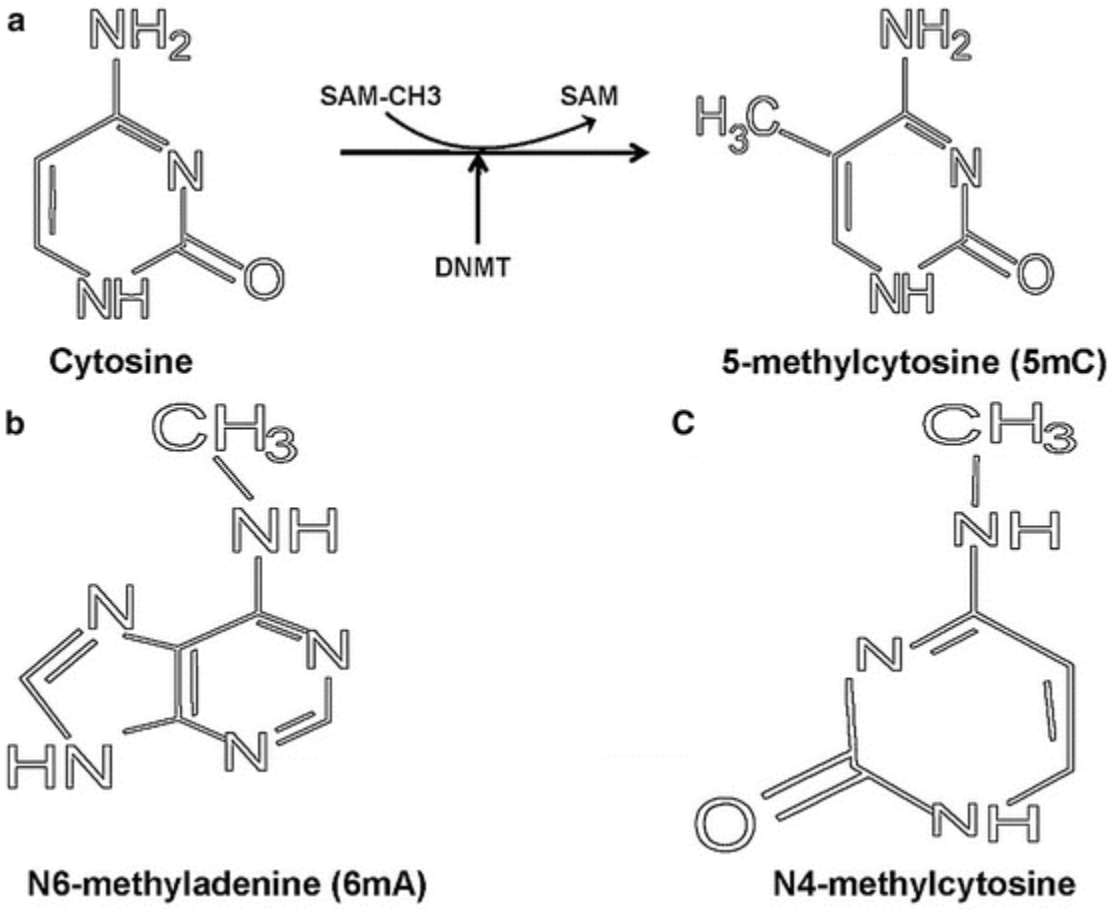


Figure 12: a) Representation of the reaction for producing c5-cytosine methylation, b) N6-methyladenine and c) N4-methylcytosine.

C5-MTs share a set of conserved motifs, nine of which are common to N4 -cytosine and N6 adenine methyltransferases (motifs I-VIII and X) (Bujnicki 2002) (Figure 13).

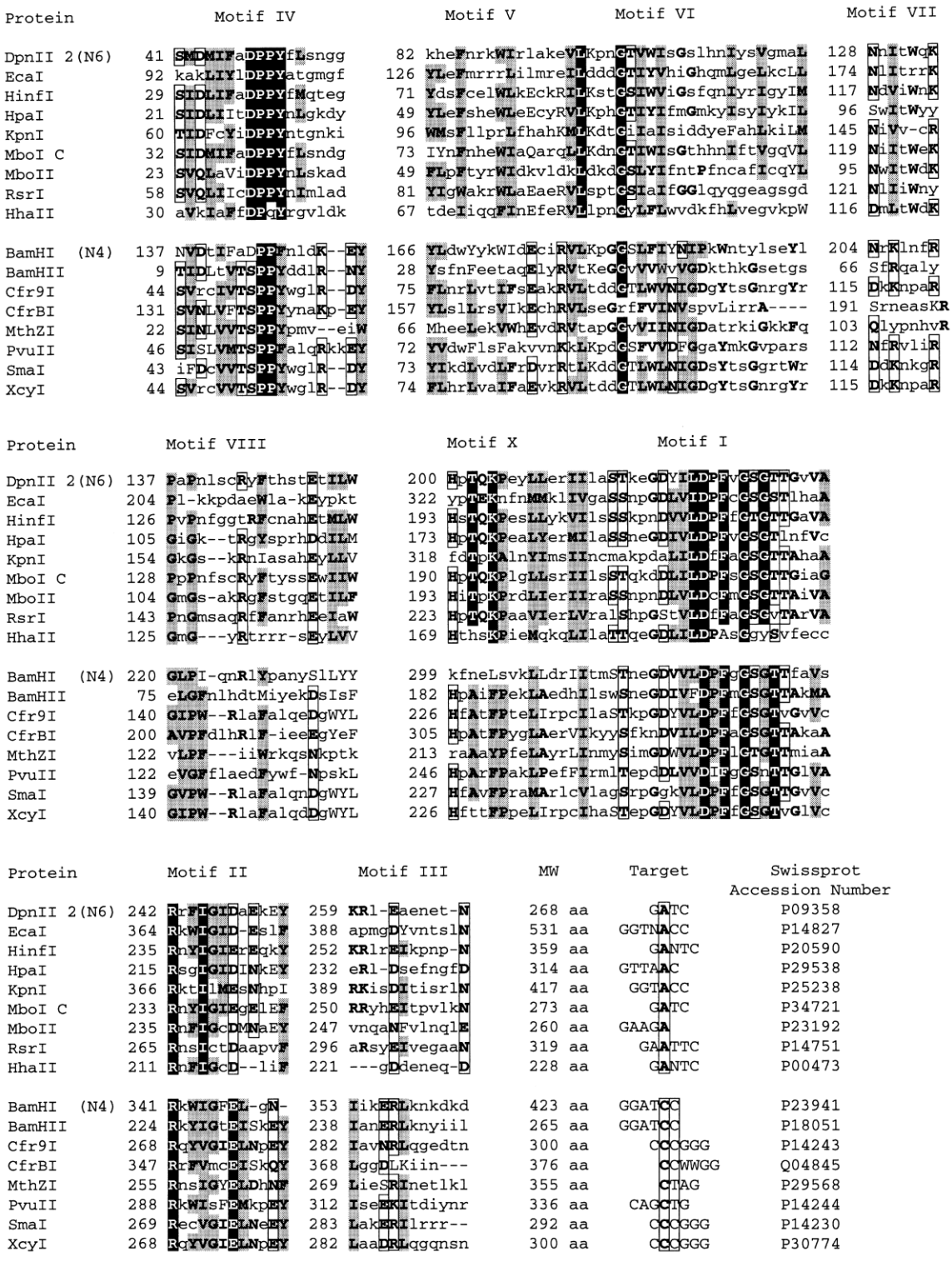


Figure 13: Nine conserved motifs among N6-adenine and N4-cytosine MTases. Invariant amino acids within a group are shown as white letters against a black background, conserved hydrophobic positions are indicated by bold letters on a shaded background, and conserved polar or charged positions by bold letters within a box. Lesser degrees of conservation are shown, in decreasing order, by bold and uppercase letters, while non-conserved positions are shown as lowercase letters. A (-) indicates a deletion relative to other sequences (Malone, Blumenthal, and Cheng 1995).

The most conserved motif among all methyltransferases is IV, which consists of the conserved residues PP(Y/F) in N-MTs and (D/P)PC in C-MTs (Timinskas, Butkus, and Janulaitis 1995). Motif IV includes the active site of MTs and in particular the proline residue participate in methylation (Bheemanaik, Reddy, and Rao 2006). Motif I is also conserved among MTs and it comprises of FxGxG which is found in AdoMet binding domain (Bheemanaik, Reddy, and Rao 2006).

Exocyclic MTs are subdivided into 6 groups according to the possible linear arrangements of 3 modules: a) AdoMet binding domain (motif I), b) active site domain (motif IV) and c) the variable target recognition domain (TRD). These groups are named α,β,γ,ζ and hypothetical δ and ε (Bujnicki 2002). The majority of exocyclic MTs are members of α, β and γ subclasses and only one MT, M.BssHI, belongs to ζ subgroup (Bheemanaik, Reddy, and Rao 2006) (Figure 14).

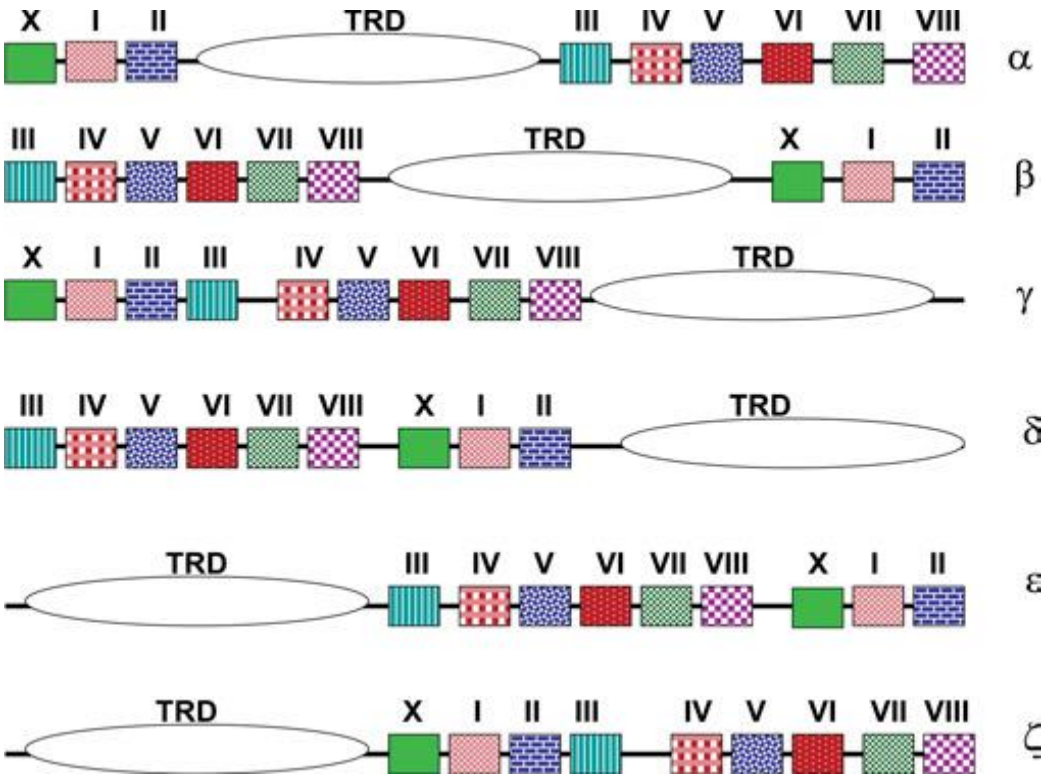


Figure 14: Classified exocyclic MTs in subgroups according to the arrangement of 3 modules:1) motif I, 2) motif IV and 3) TRD (Bheemanaik, Reddy, and Rao 2006).

1.4 The challenge of REs cloning

Cloning and overexpression of a RE to a new strain faces all problems of expressing the toxin component of a Toxin/Antitoxin system (TA). In numerous cases where TA systems were studied experimentally, cloning of the toxin was nearly impossible in the absence of the cognate antitoxin. That's also true with the RM systems. In order to clone and express a specific RE the DNA of the host cell should already be protected by the respective MT.

Usually the two genes, for RE and MT, are spatially linked within the genomic DNA and tandemly organized. Three arrangements are possible: a) in parallel, b) convergent and c) divergent (Figure 15).

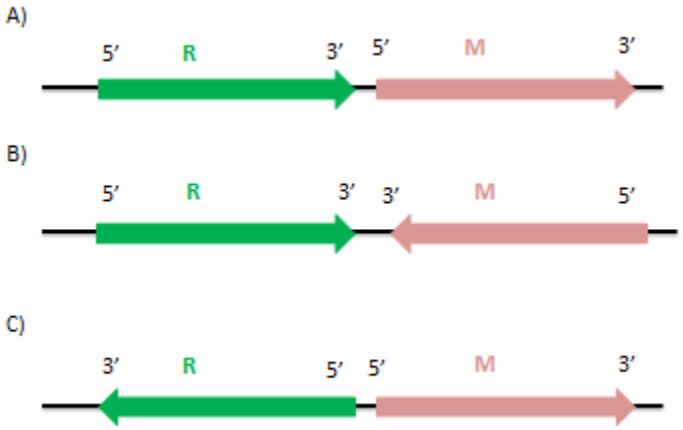


Figure 15: R-M genes A) in parallel: the 5' end of one gene following the 3' end of the other, B) convergent: the 3' ends of the genes in proximity and C) divergent: the 5' ends of the genes are near one another

RM systems can be cloned by two approaches: a.) the one-step, b.) the two steps. The first approach concerns cloning of a DNA fragment containing both genes. This fragment includes also the regulatory elements for controlling the expression of the two genes. However there are cases that this approach cannot be applied. This has to do with the fact that host genomic DNA is not fully protected by MT action either due to lower activity of the MT or due to the fast action of the respective RE. In these cases MT is cloned in a selected vector and transformed in the target strain (Alfred Pingoud, Wilson, and Wende 2014). The resulting recombinant strain is then transformed with a plasmid containing the respective RE gene or RE and MT genes. HhaII, EcoRII, EcoRI, PstI, PaeRI, EcoRV, PvuII and BsuRI are examples that have been cloned with the one step approach, while BamHI, DdeI and BglII were recombinantly overexpressed by the two steps procedure.

RM systems may include an extra protein, the C-protein. It is the controller (C-protein) that regulates the transcription so as to prevent the cell from premature expression of RE(Alfred

Pingoud, Wilson, and Wende 2014). For example, disruption of the gene BamHIC in RM system cause an increase in modification and a decrease in restriction expression.

Another challenge concerns the strains in which the RM system should be inserted. RM systems can cause problems when a vector containing DNA from another organism is introduced into a bacterial host. Some *E.coli* strains cannot survive under certain DNA methylation. Thereby they cannot hold the respective RM systems. This intolerance exists due to three endogenous *E.coli* systems, McrA (modified cytosine restriction), McrBC (modified cytosine restriction) and Mrr (modified adenine recognition and restriction). Strains such as HB101 and its derivative RR1, both defective in McrBC and Mrr proteins are widely used for cloning RM systems(Alfred Pingoud, Wilson, and Wende 2014).

2. Aim and Significance

The goal of this study is to clone and overexpress the CspAI RM system.

Whole genome sequencing of *Corynebacterium sp*. strain will provide the first view of:

- The R.CspAI amino acid sequence. Comparative analysis with other type II REs and especially with R.AgeI will support data on conserved motifs and mechanism of action
- The M.CspAI amino acid sequence. The identification of conserved motifs will help us to identify the classification of this MT

Both genes will be cloned, expressed and biochemically characterized in order to reveal their mode action.

Of great importance will be to reveal the crystal structure of R.CspAI. This will give the opportunity to compare for the first time the crystal structures of two isoschizomer REs with different sensitivities on the methylation status of target sequence. AgeI and CspAI recognize and cut the sequence 5'A/CCGGT3'. When the first C is methylated, both enzymes are blocked, while when the second C is methylated, CspAI is active whereas AgeI is blocked. It will be really interesting to compare the crystal structures of these two enzymes in order to identify the critical structural components regarding methylation sensitivity. This data will help future studies on rational design of DNA modifying enzymes.

3. Materials and Methods

3.1 Whole genome sequencing

3.1.1 Genomic DNA preparation

Total genomic DNA *of Corynebacterium species* was sequenced. 2µl of *Corynebacterium species* glycerol stock (property of Minotech Biotechnology) were plated at enhanced LB-agar plate (10 gr Yeast extract, 16 gr Tryptone, 5 gr NaCl, pH=7, 15gr agar/ lt) for 48 h at 28^{o}C. A single colony was diluted in 10ml liquid enhanced LB and incubated for 10 h at 28^{o}C until the OD_{600} reached 0.5 (vigorous shaking). Genomic DNA of *Corynebacterium* was extracted from 2 x 10^{9} cells following the protocol for the gram-positive bacteria of the PureLink Genomic DNA mini kit (Invitrogen).

3.1.2 DNA sequencing (by Genomics Facility, IMBB)

After quantification using Qubit (ThermoFisher Scientific), a library was prepared using the Ion Xpress Plus gDNA Fragment Library Kit (ThermoFisher Scientific) and purified using the Agencourt AmPure XP kit (Beckman Coulter). 100 ng sheared DNA was subsequently modified, through adapter and barcode ligation, and nick repair followed. Adapter-ligated and nick-translated DNA was purified and size selected with e-gel (Invitrogen). The target peak size of the amplified libraries was ~270 bp. The size selected library was prepared, using the Ion HI-Q OT2 200 kit (ThermoFisher Scientific). Sequencing was performed via the Ion Torrent semiconductor technology, using the Ion PI Hi-Q sequencing 200 kit and an Ion PI Chip (ThermoFisher Scientific). The machine is Ion S5 XL System (Thermofischer).

3.1.3 Annotation

Due to the absence of a reference genome, *de novo* annotation was followed in order to determine the coding regions. *De novo* sequencing was performed so as to identify the spatially linked coding regions of the restriction enzyme and the methylase.

3.2 Cloning of M.CspAI and R.CspAI

3.2.1 Chemically Competent cells

A fresh streaking from a glycerol stock (20% glycerol) was performed and a colony was cultured in 10ml liquid LB (5 gr NaCl, 5gr Yeast Extract and 10 gr Tryptone, pH=7) for 16h at 37 ^{o} C. The pre-culture was used to inoculate 100ml LB in a ratio of 1:200 until the OD_{600} reached 0.5. Subsequently, the culture was let in ice for 10 minutes and the cells were collected by centrifugation for 13 minutes at 2000 g at 4^{o}C. The cell paste was resuspended (mild shaking, in ice) with 10 ml sterile TFBI (30mM CH_3COOK, 50mM $MnCl_2$, 100mM KCl, 10mM $CaCl_2$ and 15% glycerol). The cells were collected by centrifugation for 8 minutes at 2000 g at 4^{o}C. The paste

was resuspended (mild shaking, in ice) with 2 ml TFBII (75mM $CaCl_2$, 10mM KCl, 10mM MOPS and 15% glycerol). Finally the solution was aliquoted (50μl or 100μl/aliquot) and frozen immediately in a dry ice-ethanol bath or in liquid nitrogen and stored at – 80°C. All steps were carried out under sterile conditions.

3.2.2 Cloning of M.CspAI

The pACYC184 vector (supplied by New England Biolabs) was selected for cloning the methylase gene. A set of primers were designed including the appropriate restriction site sequences. Four forward primers were designed including a BamHI restriction site, and a common reverse primer was also designed including a SalI restriction site (Table 4). Forward 1 primer contains only the restriction site of BamHI and the sequence of the gene. Forward 2 primer contains 43 nucleotides before the gene which are the same as the region 1875-1917 of pLysE vector (EMD Biosciences). Forward 3 includes only RBS and an NdeI restriction site. Forward 4 was designed based on the RBS Library Calculator (DE NOVO DNA) setting as parameters the pACYC184 sequence up to BamHI site (actatcgactacgcgatcatggcgaccacacccgtcctgtggatcc) and the coding sequence of M.CspAI gene.

	Name	Sequence
Forward 1	F1	5'-CGCggatccATGCAGGGGGACTGTCTGG-3'
Forward 2	F2	5'-CGCggatccgggtccccttgatagattaaaaaggaaaggaggaaagaaataATGCAGGGGGACTGTCTGG-3'
Forward 3	F3	5'-CGCggatccAAGGAGATATACATATGCAGGGGGACTGTCTGG-3'
Forward 4	F4	5'-CGCggatccGTAGAACGTCACATAAAGGGACATAATATGCAGGGGGACTGTCTGG-3'
Reverse	R	5'-ACGCgtcgacTTAAGAGACTGGCTTGTCGTCTAAGATC-3'

Table 4: The designed primers are shown in the table. With red are depicted the restriction sites (BamHI and SalI). In F3 primer in blue is depicted the NdeI restriction site and the RBS site is underlined. In F4 primer, in green are shown the sequence from RBS Library Calculator.

Four PCR reactions were performed (Table 5)

Reaction Components	Reaction Conditions
10μl 5x Minopol Buffer	1. 98 ° C for 30 seconds
0.2mM dNTPs	2. 98 ° C for 10 seconds
0.5μM F1/F2/F3/F4 primer	3. 65 ° C for 30 seconds
0.5μM R primer	4. 72 ° C for 26 seconds
10ng DNA Template	Repeat steps 2-4 for 30 times
0.5 Unit Minopol Polymerase	5. 72 ° C for 10 minutes
ddH_2O to a final volume of 50 μl	6. Retention at 4 ° C

Table 5: PCR components and conditions for amplification of M.CspAI gene.

After each PCR reaction, the subsequent product was visualized in 1% agarose gel with ethidium bromide straining, gel extracted using the Monarch DNA Gel Extraction Kit (New England Biolabs) and quantified by a Nanodrop spectrophotometer (Thermofischer).

Digestion reactions of the pACYC184 plasmid and the four amplified DNA genes were performed as follows: 1X Unique Buffer, 12 units of BamHI, BSA 100μg/ml and 2.5μg of plasmid DNA and 1μg of amplified DNA genes, respectively, to a final volume of 50μl. The conditions of the

digestion were 3h at 37°C. After a clean-up assay with Monarch PCR and DNA Clean up Kit (New England Biolabs) a second digestion reaction followed with 50 units of SalI while the reaction conditions were the same as above. After a gel extraction assay using Monarch DNA Gel Extraction Kit , 140ng of the plasmid DNA was ligated with the four DNA genes in separate reactions in a ratio 1:3 (plasmid:insert). The reaction conditions were as follows: 1X Ligase Buffer, 10mM ATP, 2.5 units of T4 DNA Ligase and the appropriate volumes of plasmid DNA and insert in an overnight incubation at 16°C. A ligation reaction only with digested pACYC184 was performed as negative control.

DH5a chemically competent cells were transformed with 3μl of the ligated product based on a heat-shock transformation protocol. The cells were plated on LB-Agar plates with 34 μg/ml Chloramphenicol. Respective colonies were tested by colony PCR. The positive colonies were cultured in 5ml liquid LB containing 34 μg/ml Chloramphenicol at 37°C for 16h and the plasmid purification was performed using the Plasmid DNA Purification kit (Macherey-Nagel). A digest with NdeI enzyme was performed for verification of the desired constructs. REs were supplied from MINOTECH Biotechnology and New England Biolabs.

3.2.3 pGEM T-easy cloning of M.CspAI

The methylase gene was cloned using the pGEM T-easy cloning kit (Promega). The gene was amplified by PCR reactions (Table 6) using the upstream primers F1, F2, F3 and F4 and the downstream primer R (Table 4).

Reaction Components	Reaction Conditions
5μl 10x Taq pol.Buffer with $MgCl_2$	1. 94 ° C for 2 minutes
0.2mM dNTPs	2. 94 ° C for 45 seconds
0.25μM F1/F2/F3/F4	3. 60 ° C for 30 seconds
0.25μM R	4. 72 ° C for 55 seconds
10ng DNA Template	Repeat steps 2-4 for 30 times
5 Units Taq Polymerase	5. 72 ° C for 10 minutes
ddH_2O to a final volume of 50 μl	6. Retention at 4 ° C

Table 6: Four PCR reactions for amplification of the methylase gene with Taq DNA polymerase (Minotech Biotechnology).

Each PCR product was purified, following the gel extraction assay, and ligated to pGEM vector. The reaction mixture consisted of 1x Buffer (supplied by kit), 50ng pGEM, 3 Units of T4 DNA ligase (supplied by kit) and PCR product to a final volume of 10μl. The reaction was incubated for 16h at 4°C. DH5a chemically competent cells were transformed with 5μl of the ligated product based on a heat-shock transformation protocol. The cells were plated on LB-Agar plates with 100μg/ml Ampicillin, 0.2μg/ml IPTG and 64μg/ml X-gal. After blue/white selection, colony PCR was performed on white colonies. The positive colonies were cultured in 5ml liquid LB containing 100μg/ml Ampicillin at 37°C for 16h and the plasmid purification was performed using the Plasmid DNA Purification (Macherey-Nagel). A digest with NdeI was performed for verification of the desired constructs. The resulting plasmid constructs were sequenced

(Macrogen) for verification. The sequencing results were analyzed using the Vector NTI software package (INVITROGEN).

Two constructs were selected, pGEM+F1 and pGEM+F4. Both were digested sequential with BamHI-HF (New England Biolabs) and SalI (Minotech Biotechnology). The same sequential digestion was followed for vector pACYC184. Ligation reaction was performed in ratio 1:5 for F1:pACYC184 and 1:10 for F4:pACYC184 and 3µl of each ligation reaction were transformed in DH5a chemically competent cells and plated on LB-Agar plates with 34µg/ml Chloramphenicol. The colonies were verified by colony PCR and respective purified plasmids were tested with HindIII/NdeI digest. The positive constructs were verified by sequencing and resulted to plasmid pACYC184-F1 M.CspAI and pACYC184-F4 M.CspAI.

3.2.4 Cloning of RCspAI

The RCspAI gene was cloned in pET16b vector with Ampicillin resistant and a His-tag in the N-terminus site. A forward primer was designed including a BamHI restriction site and the reverse primer a SlaI site (Table 7). The forward primer does not include ATG start codon of RCspAI since it already exists prior to N-terminus His-tag.

	Name	Sequence
Forward pET16b	F.HN	5'- cctcgagTTGATCGAAGTATCATCGCAG-3'
Reverse pET16b	R.HN	5'- cggatccTTAACTCATGAACTCTTGGTAGGC-3'

Table 7: Primers for cloning RCspAI in pET16b vector. With Red are depicted the restriction sites.

PCR reaction was performed as in Table 8.

Reaction Components	Reaction Conditions
10µl 5x Minopol Buffer	1. 98 ° C for 30 seconds
0.2mM dNTPs	2. 98 ° C for 10 seconds
0.5µM F.HN	3. 60 ° C for 30 seconds
0.5µM R.HN	4. 72 ° C for 40 seconds
10ng DNA Template	Repeat steps 2-4 for 30 times
0.5 Unit Minopol Polymerase	5. 72 ° C for 10 minutes
ddH_2O to a final volume of 50 µl	6. Retention at 4 ° C

Table 8: PCR reaction for amplification of the R.CspAI gene.

The amplified PCR product was purified using the Monarch DNA Gel Extraction Kit.

2µg pET16b vector and 1µg amplified gene R.CspAI were digested with BamHI and SlaI as follows: 1x K-Buffer, 12 units of BamHI and 12 units SlaI to a final volume of 50µl. The conditions of the digestion were 2h and 1.5h, respectively, at 37^{o}C. After a clean-up assay with Monarch PCR and DNA Clean up Kit for the digested R.CspAI gene and Monarch DNA Gel Extraction Kit for digested pET16b, 100ng of the plasmid were ligated in a ratio 1:3 (plasmid:insert). The reaction conditions were as follows: 1X Ligase Buffer, 10mM ATP, 2.5 units of T4 DNA Ligase and the

appropriate volumes of plasmid DNA and insert in an overnight incubation at 16^{o}C .A ligation reaction only with digested pET16b was performed as negative control. DH5a, RR1 and HB101 (Supplementary data, Table 1) chemically competent cells were transformed with 2μl of the ligated product based on a heat-shock transformation protocol. The cells were plated on LB-Agar plates containing 100 μg/ml Ampicillin. No colony appeared after 24h incubation.

Primers for cloning R.CspAI in pET24c were also designed so as that R.CspAI contains a His-tag at its C-terminus end (Table 9). The cloning was unsuccessful.

	Name	Sequence
Forward pET24c	F.HC	5' ATGTTGATCGAAGTATCATCG -3'
Reverse pET24c	R.HC	5'- cctcgagACTCATGAACTCTTGGTAGGC -3'

Table 9: Primers for cloning R.CspAI in pET24c vector.

3.2.5 Genome assembly Neb builder kit for cloning R.CspAI

The gene of R.CspAI was also cloned with the NEBuilder HiFi DNA Assembly Cloning Kit (New England Biolabs). This kit is a modified version of Gibson Assembly cloning method (10). In particular, the fragments, vector pET16b and the R.CspAI gene, were amplified by PCR using the primers presented on Table 10. Subsequently, a ligation reaction was performed according to the instructions of the protocol for two DNA fragments. The reaction is consisted of three different enzymes: an exonuclease which creates 3' prominent ends so as to facilitate the hybridization of the complementary fragments, a polymerase which leads to a double-stranded DNA and a ligase which binds the fragments that have been produced.

	Name	Sequence
Forward for RCspAI	RCspAI HN F	5'-GCCATATCGAAGGTCGTCATATGTTGATCGAAGTATCATCGC-3'
Reverse for CspAI	RCspAI HN R	5'-GGGCTTTGTTAGCAGCCGGATCCTTAACTCATGAACTCTTGGTAG-3'
Forward for pET16b	pET16b F	5'-GGATCCGGCTGCTAACAAAG-3'
Reverse for pET16b	pET16b R	5'-CATATGACGACCTTCGATATGGC-3'

Table 10: Primers were designed for both vector and insert according to the NEBuilder kit.

F' I^{q} E.coli competent cells were transformed with 2μl of the reaction and plated on LB-agar plates containing 100μg/ml Ampicillin. Respective colonies were tested by colony PCR and the positive ones were grown in 5ml liquid LB with 100μg/ml Ampicillin for 16h at 37^{o}C. The desirable plasmid constructs were purified using Plasmid DNA Purification kit from Macherey-Nagel. A digest with HindIII (Minotech Biotechnology) was performed for verification of the desired construct. The positive constructs were verified by sequencing and resulted to plasmids pET16b-R.CspAI.

3.2.6 Site directed mutagenesis

This protocol is used to make point mutations, replace amino acids, and delete or insert single or multiple adjacent amino acids. The method is performed using a high fidelity DNA polymerase for mutagenic primer-directed replication of both plasmid strands with the highest fidelity. The basic procedure utilizes a supercoiled double-stranded DNA (dsDNA) vector with an insert of interest and two synthetic oligonucleotide primers, both containing the desired mutation. The oligonucleotide primers, each complementary to opposite strands of the vector, are extended during temperature cycling by the high fidelity DNA polymerase, without primer displacement. Extension of the oligonucleotide primers generates a mutated plasmid containing staggered nicks. Following temperature cycling, the product is treated with Dpn I. The Dpn I endonuclease (target sequence: 5′-Gm6ATC-3′) is specific for methylated and hemimethylated DNA and is used to digest the parental DNA template and to select for mutation-containing synthesized DNA. (DNA isolated from almost all *E. coli* strains is dam methylated and therefore susceptible to Dpn I digestion.) The nicked vector DNA containing the desired mutations is then transformed into DH5α cells (Figure 16).

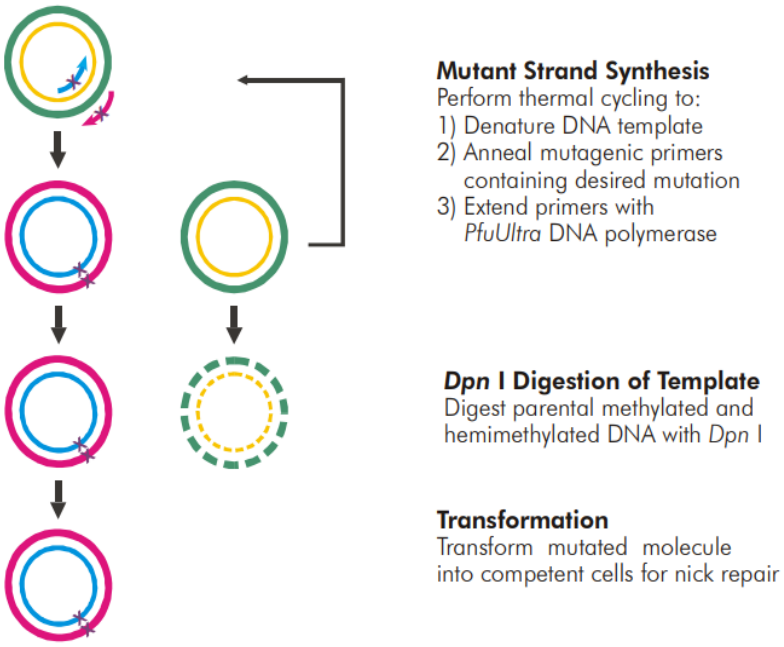


Figure 16: Overview of site-directed mutagenesis method.

The primers used to construct the wild type gene of methylase are presented below (Table 11).

	Name	Sequence
Forward M.CspAI F1	MCspAI F1 FGA	TTTATTGGCC**A**AAAGTGAAAAGCATG
Reverse M.CspAI F1	MCspAI F1 RGA	CTTTTCACTTT**T**GGCCAATAAATAGATCG
Forward M.CspAI F4	MCspAI F4 FGA	GACTTGTGCG**A**TGCTGGATGGATATTCC
Reverse M.CspAI F4	MCspAI F4 RGA	CCATCCAGCA**T**CGCACAAGTCAATCAC

Table 11: Primers for the wild type of M.CspAI. In bold are depicted the desired mutation.

Reaction Components	Reaction Conditions
10μl 5x Minopol Buffer	1. 98 ° C for 30 seconds
0.2mM dNTPs	2. 98 ° C for 10 seconds
0.5μM MCspAI F1 FGA/ MCspAI F4 FGA	3. 62° C /67 ° C for 30 seconds
0.5μM MCspAI F1 RGA/ MCspAI F4 RGA	4. 72 ° C for 1min and 12 seconds
10ng/2ng/1ng DNA Template	Repeat steps 2-4 for 18 times
0.5 Unit Minopol Polymerase	5. 72 ° C for 10 minutes
ddH_2O to a final volume of 50 μl	6. Retention at 4 ° C

Table 12: PCR reaction for site directed mutagenesis. 3 different concentrations (10ng, 2ng and 1ng) of DNA template (pACYC184-F1 M.CspAI and pACYC184-F4 M.CspAI) were amplified in two annealing Temperatures (62° C and 67 ° C).

The PCR products (10μl of 50μl) were visualized in 1 % agarose gel with ethidium bromide straining, and the linear product was treated by 16 units DpnI (supplied by New England Biolabs) for 1h at 37^{o}C. 10μl of each reaction was used to run in 1% agarose gel and 5μl of each reaction were transformed in DH5a cells and plated on LB agar plates with 34μg/ml Chloramphenicol. Colony PCR was performed to identify the positives colonies, which, subsequently, were cultured in liquid LB and 34 μg/ml Chloramphenicol for 16h at 37^{o}C vigorous shaking. The positive constructs were verified by sequencing and resulted to plasmids pACYC184-wt F1 M.CspAI and pACYC18-wt F4 M.CspAI.

3.2.7 Colony PCR

After each cloning step, the desired colonies were diluted in 20μl sterile ddH_2O. In order to confirm that the colonies contain the desired construct a PCR reaction was performed using the Taq polymerase (Minotech Biotechnology) and the appropriate primers. The PCR schedule is shown on Table 13

Reaction Components	Reaction Conditions
5μl 10x Taq pol.Buffer with $MgCl_2$	1. 94°C for 2 minutes
0.2mM dNTPs	2. 94°C for 45 seconds
0.5μM Forward primer	3. Primers Tm for 30 seconds
0.5μM Reverse primer	4. 72°C for 1min/kb
5μl of diluted colony	Repeat steps 2-4 for 30 times
5 Units Taq Polymerase	5. 72 ° C for 10 minutes
ddH_2O to a final volume of 50 μl	6. Retention at 4 ° C

Table 13: PCR components and conditions for colony PCR

3.3 Overexpression of R.CspAI in E.coli cells

3.3.1 Transformation

The chemically competent cells were transfer from – 80^{o}C to 4^{o}C for 5 minutes. Under sterile conditions the appropriate amount of DNA (less than 200ng, less than 10% of competent cell volume) was added to the cells and let in ice for 30 minutes. A heat-shock was performed incubating the cells at 42^{o}C for 1.5 minute and immediately transfer in ice for 2 minutes. The cells were incubated with 900μl SOC medium (20 gr Tryptone, 5gr Yeast extract, 10mM NaCL, 2.5mM KCl, 20mM $MgSO_4$ and 0.4% glycerol, pH=7) at 37^{o}C for 1h shaking. The cells were collected by centrifugation for 5 minutes at 1100 g, RT and plated on LB-agar plates using the appropriate antibiotics. Plates were incubated at 37^{o}C for 16h

3.3.2 Transformation of pACYC184-M.CspAI and pET16b-R.CspAI plasmid constructs

Two methods were followed for inserting two plasmid constructs in one cell. In the first case, T7 express *E.coli* competent cells (New England Biolabs) were transformed simultaneously with pET16b-RCspAI and pACYC184-wt M.CspAI. 10ng and 50 ng of each plasmid construct were imported in competent cells following a heat- shock transformation protocol and the transformed cells were plated on LB-agar plates containing 100 μg/ml Ampicillin, 34 μg/ml Chloramphenicol and 1% glucose. A fresh colony was chosen for the inoculation of a LB pre-culture, in the presence of the above mentioned concentrations of the two antibiotics and incubated overnight (16 hours) at 37^{o}C.

In the second case, T7 express *E.coli* competent cells, containing the pACYC184-wt M.CspAI plasmid, were transformed with 10ng and 20ng pET16b-R.CspAI plasmid construct. After a heat-shock transformation protocol, the transformed cells were plated on LB-agar plates with 100μg/ml Ampicillin, 34 μg/ml Chloramphenicol and 1% glucose. The colonies were tested by colony PCR for R.CspAI gene. The positives colonies were cultured in 5ml liquid LB containing 100 μg/ml Ampicillin, 34 μg/ml Chloramphenicol and 1% glucose at 37^{o}C for 16h under vigorous shaking.

3.3.3 R.CspAI overexpression

The pre-culture was used to inoculate 50 ml of LB in a ration of 1:20, in the presence of Ampicillin, Chloramphenicol and 1% glucose. The inoculated cultures were grown:

a.) at 37^{o}C until OD_{600} reached 0.5 and were centrifuged at 2330 g for 10 minutes at 4^{o}C. The cell paste was resuspended in 50ml LB with Ampicillin, Chloramphenicol and the induction reagent IPTG (IsoPropyl- beta-D-thioGalactopyranoside) at a 0.5mM final concentration. Cultivation continued for another 3 hours at 37^{o}C under shaking. The cells were collected by centrifugation of the cultures at 2330 g for 10 minutes at 4^{o}C. The cell pastes were stored at - 20^{o}C until use.

b.) at 37°C until OD_{600} reached 1-1.5 and the process as above was followed. After the addition of IPTG, cultivation was continued for 16h (overnight) at 16°C under shaking. The collection method was the same as above.

As controls, T7 express cells and T7 express cells with pACYC184-wt M.CspAI (F1 and F4) plasmid constructs were pre-cultured in 5 ml LB and LB-Chloramphenicol respectively for 16h at 37°C. The pre-cultures were used to inoculate 50ml LB-chloramphenicol in a ratio 1:20. The cultures were grown until the OD_{600} reached 1. IPTG at a final concentration 0.5mM was added to each culture and the cultivation continued for 16h at 16°C shaking. The collection method was same as above.

For a large scale culture a fresh transformation was performed and a colony was selected and cultured in 50ml LB supplemented with 100μg/ml Ampicillin, 34 μg/ml Chloramphenicol and 1% glucose at 37°C for 16h. The pre-culture was used to inoculate 1 liter LB containing the above antibiotics and 1% glucose at a ratio 1:80. The culture was grown until the OD_{600} reached 0.5 and the cultures were centrifuged at 5000g, at 20°C for 12 minutes. The cell pastes were respundended in 1 liter LB- Ampicillin-Chloramphenicol and IPTG at a final concentration 0.5mM. The cultivation continued for 3h at 37°C shaking and the cells were collected by centrifugation at 5000g, at 4°C for 12 minutes. Cell pastes were stored at -20°C until use.

As controls, T7 express cells containing pACYC184-wt M.CspAI (F1 or F4) were pre-cultured in 50 ml LB supplemented with 34μg/ml Chloramphenicol at 37°C shaking for 16h. The pre-culture was used to inoculate 1 liter LB-Chloramphenicol at a ratio 1:80. The cultures were grown at 37°C shaking until the OD_{600} reached 0.8. The collection method was the same as above.

3.4 R.CspAI Purification

3.4.1 Cell disruption

Frozen cell pastes from 50 mL cell culture (>0.1 gr) was thawed on ice and resuspended in 1 ml of a solution containing 20 mM Tris-HCl pH: 7.6 and 200 mM NaCl. Cells were disrupted by sonication (5x30 sec sonication steps, with 30 sec resting intervals on ice) and the lysate was centrifuged at 20.800 g, at 4°C for 20 minutes. The cytoplasmic supernatant was collected for analysis by Western blot.

In addition, frozen cell pastes from 1 liter cell culture (approximately 1gr) was thawed on ice and resuspended in equal volume of Lysis Buffer containing 20 mM Tris-HCl pH: 7.6, 200 mM NaCl and 5mM beta-mercaptoethanol. Cells' disruption was performed by sonication (10x30 sec sonication steps, with 30 sec resting intervals on ice) and the supernatant was collected by centrifugation at 20.800 g, at 4°C for 30 minutes. The cytoplasmic supernatant was collected for loading to the chromatography column.

3.4.2 Affinity Chromatography with Ni-NTA column

A chromatography column containing 0.5ml (1 column volume) Ni-NTA beads were equilibrated and loaded with the diluted supernatant as follows. 2 CV (column volume) 0.1 M NaOH was added and mild shaking was performed for 15 minutes at 10^{o}C. After centrifugation (2 minutes, RT, at 270 g), the supernatant was discarded and addition of 2 CV 2M NaCl was followed (5 minutes, RT, mild shaking). A centrifugation was followed and the column was equilibrated with 20 CV Lysis buffer (mild shaking and centrifugation). After the removal of supernatant the column was loaded with the diluted sample in ratio 1:4 in Lysis Buffer and incubated under mild shaking for 2h at 10^{o}C. The unbound proteins were removed by 5 CV Lysis Buffer and the bound protein was collected by two elution steps, step1: 1 CV of elution solution (20mM Tris-HCl pH=7.6, 200mM NaCl, 5mM beta-mercaptoethanol and 600mM Imidazole) and step2: 2 CV of elution solution. After a mild shaking for 5 minutes and centrifugation, the supernatant was collected and the fractions were analyzed by SDS-PAGE and Western blot in order to validate the overexpression of R.CspAI.

3.5 Western blot

The samples were transferred from a 12.5% SDS-PAGE to a nitrocellulose membrane using a Transfer Solution (1x Electrophoresis Buffer and 20% Methanol). The transfer was performed for 1h and 15 minutes at 4^{o}C, at 138 Volt. Subsequently, 50 ml blocking solution (1% TTBS and 5% dry milk powder) was added for 1h (RT, mild shaking). After three wash steps with 10 ml 1x TTBS (20mM Tris-HCl pH=7.5, 150mM NaCl and 0.3% Tween-20) 50 ml mouse anti-His Antibody was added in dilution 1:200 for 16h shaking at 4^{o}C. After three wash steps, 50ml anti-Mouse Antibody was added (1:25000 dilution) and incubated for 1h RT, mild shaking. 1 ml of each ECL reagents (Thermofischer) was mixed and let for 5 minutes, RT. After four washes of the membrane (three with 1xTTBS and one with 1x TBS) the ECL mix was loaded on the membrane on the protein side and incubated for 1 minute without shaking. The images were acquired by Chemi-Doc (Biorad) after 30 seconds exposure time.

3.6 Mass spectrometry (PRoFI, IMBB)

A sample of R.CspAI supplied by Minotech biotechnology was analyzed in a keratin free 12.5% SDS-PAGE .Two bands were excited from the gel and an in-gel tryptic digestion protocol was followed (Protocols are available in PRoFI, IMBB, FORTH). Each sample was lyophilized by Speed-Vacuum and the dried peptides were resuspended by 10µl 0.5% formic acid. Each sample was measured with nLC-ESI-MS/MS (LTQ-Orbitrap XL from ThermoFischer Scientific) and the .raw data, collected by nLC-MS/MS, were loaded in Proteome Discoverer 1.3, running Mascot Matrix.

3.7 Computational tools

Bioinformatics analysis (plasmid maps, sequencing analysis, align x) were performed by Vector NTI software package (INVITROGEN) and https://pfam.xfam.org/ for protein family prediction.

Primers Tm were calculated by http://tmcalculator.neb.com/#!/main and ligation reagents by http://2011.igem.org/Team:UT_Dallas/ligation.

4. Results

4.1 Identification of CspAI RM system sequence

In order to heterologously express the components of the CspAI RM system we need to know the respective sequences of the target genes. Two different methodologies were followed in parallel: the proteomic and the genomic approach (Figure 17).

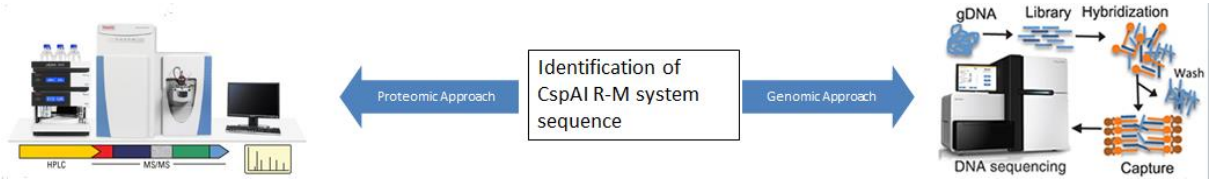
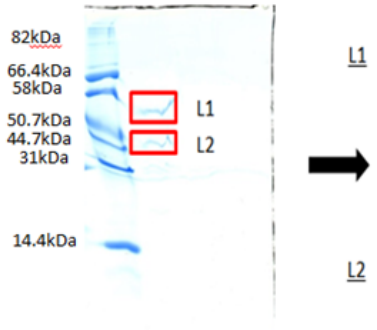


Figure 17: For the identification of the R.CspAI and M.CspAI sequence two approaches were followed, the proteomic and the genomic.

R.CspAI enzyme solution was available from Minotech Biotechnology, so the proteomic analysis used it as the starting material. A sample of R.CspAI protein was analyzed in SDS-PAGE and two protein bands appeared after coomasie staining (Figure 18). Since there was no information regarding the exact size, both SDS-PAGE bands were analyzed. After proteolysis and subsequent nLC-ESI-MS/MS analysis, results were further analyzed by Proteome discoverer, running Mascot Matrix searching in all *Corynebacterium* proteomes for matching proteins. However none of them had 100% coverage with a RE or MT of *Corynebacterium* (Figure 18).



L1

Accession	Description	Score	Coverage
A0A0S2T4F7	Glutamate--tRNA ligase OS=Corynebacterium glutamicum GN=gltX PE=3 SV=1 - [A0A0S2T4F7_CORGT]	3609,14	67,94
Q8FPU9	Glutamate--tRNA ligase OS=Corynebacterium efficiens (strain DSM 44549 / YS-314 / AJ 12310 / JCM 11189 / NBRC 100395) GN=gltX PE=3 SV=2 - [SYE_COREF]	1103,18	30,43

L2

Accession	Description	Score	Coverage
A0A072ZMH0	Preprotein translocase subunit SecA OS=Corynebacterium glutamicum ATCC 14067 GN=KIQ_002485 PE=4 SV=1 - [A0A072ZMH0_CORGT]	159,47	24,34
A0A072ZM33	Uncharacterized protein OS=Corynebacterium glutamicum ATCC 14067 GN=KIQ_001920 PE=4 SV=1 - [A0A072ZM33_CORGT]	131,84	26,90

Figure 18: Since there was no information about the amino acid sequence of CspAI RM system both SDS-PAGE bands were analyzed. Lane 1 (L1) and Lane 2 (L2) were analyzed by nLC-ESI-MS/MS and the results were analyzed by Mascot matrix. The first table depicts the results from L1 and the bottom table the scores for L2.

The second approach involved whole genome sequencing of *Corynebacterium species*. Full genome assembly and subsequent full annotation has revealed a RE followed by a MT in the target genome (Figure 19a).

a)

b) **Methyltransferase gene, M.CspAI gene**

TTGGATAACATTGGTCCGTATCGGATAAATTCAGTAGTGCAGGGGGACTGTCTGGAACTT
ATTGAGGAATTGCCGAACGACAGTCTCGATGTCGTTGTGACCTCACCGCCATACTGGGGA
CAGAGAATGTCTAATGGAATGGGAGTTGAGGATGATCCTCGGGATTACCTTAAATCACTT
AATGAGATTTTTACCGCCATTCTTCCAAAGATGAAGGAGTCTGGCATTATGTGGATCAAC
ATTGGGGATTCGTACAATACACCTGTAAATTGGAGCCTTAACGATAGGAACTATTCTTCA
CTTGGTCTAAACCGGACTGGGCTTGATGCAAATAATTCTGCGTATACTAAGCCTCGTGCA
AAGCGTAGAGCATTCATTGAAAAAGAAACTCCTTGGCTGACATATGGAAACTTGCTGGCG
TTGACTTATCGCATGGTGATTGACTTATGCGATGCTGGATGGATATTCCGAGGCGAAGTT
ATCTGGCAAAAGAAGAACCCAATGCCAGAAGGTCGCGCTCGGCGTCCTCATCGTGGGCAT
GAGCCGATCTATTTATTGGCCAAAAGTGAAAAGCATGCCTTTCGGGTCAAGCCACCAGTT
AAAAGTGTGTGGACTTTCGGCAATGAAAAGATTGATGGACTTCCCCACTTTTCACGTTTT
CCGATCAAACTTCCATACCTGTGTATTGACGCCTATGGCGAGAAGGGGCAAGATGTGGTT
GTTTTTGACCCATTCTCAGGTAGTGGATCGACGGGTATTGCCGCACTAGACTTGGGTTGC
TCGTTTATAGGTTTTGAGATTGATGAAGAGCAAGTCGTGGCTTCGAATGAACGACTTCAA
CGGATCTTAGACGACAAGCCAGTCTCTTAA

Methyltransferase, M.CspAI

MDNIGPYRINSVVQGDCLELIEELPNDSLDVVVTSPPYWGQRMSNGMGVEDDPRDYLKSL
NEIFTAILPKMKESGIMWINIGDSYNTPVNWSLNDRNYSSLGLNRTGLDANNSAYTKPRA
KRRAFIEKETPWLTYGNLLALTYRMVIDLCDAGWIFRGEVIWQKKNPMPEGRARRPHRGH
EPIYLLAKSEKHAFRVKPPVKSVWTFGNEKIDGLPHFSRFPIKLPYLCIDAYGEKGQDVV
VFDPFSGSGSTGIAALDLGCSFIGFEIDEEQVVASNERLQRILDDKPVS

c) **Restriction Enzyme gene, R.CspAI gene**

GTGTTGATCGAAGTATCATCGCAGGTGCACGCTAAGCTAGGAATCATGCCTAATTTGTCT
AATGCAACGCTTGCACTTCTCAATTCTCTGAAGCAAGCTCGTGAGTTCAAAGGGATCACA
GTAGATCAAGTTGAAGAGAAATTAAAGCTTGGGCCAGGATGGATAAATGCAATTGAAAAT
GGATTAGTGACAACTGATCTTGACACCTTTATTTTCTTAGCCTCCTTGGTAGAACTCGAT
TCTAAGGAAATCTCTAAGGCGCTAGACGACAGAAGCACCGAAGATTCAATAGTCGATGTA
CTCGATAGGTCTTTCACTGCCGAACAGCAAAACAACGACTTAATTATCAAATTCAACTAC
GGTTCGCACGATGCGACCTATTTACTTGAAGGAGCCACCCTCAAGCAATTTGAAGACGTT
GTAAAACAACTAAGAGACGGCCTGGCCGTCGGAGGTTCAGCAACCCCCGACGAACAAAAT
AACCAATTCAAAGCAACTGCAGTCAAAAATTCATTTGAGCAGGCCATGAAATACTGGCCA
GATGCTAATCCCTCCGATATTTGGTATTTTCTAATTTATCGGGCCTTCTGCGATCGCTAT
AATCATCCTTCCAATGAAGCGAGATTAGCCTTTGAACAATCATGGAAGCGAACAGGCGGA
TGGGCACTAGAAATGGTTATGGTTGATCATTATGCGCCAACCCTGGCAAAGAAGGGTATC
AAACTTGAGATCGTGGATAAATTGCGAGCTACTACTATCCTAGATCGACTCAATCTCGAT
GTTCGTCTAGAAGCTGATAAAGTCGACGCATTTCTTTTTGACGAAAATGATGACTTCTTC
GGAATCATTAATGTAAAAGCTAGCTTCGCCGAACGACGAACCGATGATGTACCAATGAGT
GAAGCCCTTAACAATCGAGGCTATTACTCGGCACTGTGGACTATGGACTGTAAAGCAAAC
CCAGGTCCAAACCCATACAATAAGGGCGAATTAGGTTCGCCCGAAGGTAAACGATCAGCC
AAGCGCAAGGACATTGAAGACGAAGGGTACTTCACTGCCTGTTTTTCTTATAATCAAAAC
ACCCAGCCGACCAGCCGCAACCGTGCGACGAAAGGCCAAGTCATCATATGCGACTTCAAA
GACCCTGATGATGCATTTGTTGAGCATATCGTCGAAGCCTACCAAGAGTTCATGAGTTAA

Restriction Enzyme, R.CspAI

MLIEVSSQVHAKLGIMPNLSNATLALLNSLKQAREFKGITVDQVEEKLKLGPGWINAIEN
GLVTTDLDTFIFLASLVELDSKEISKALDDRSTEDSIVDVLDRSFTAEQQNNDLIIKFNY
GSHDATYLLEGATLKQFEDVVKQLRDGLAVGGSATPDEQNNQFKATAVKNSFEQAMKYWP
DANPSDIWYFLIYRAFCDRYNHPSNEARLAFEQSWKRTGGWALEMVMVDHYAPTLAKKGI
KLEIVDKLRATTILDRLNLDVRLEADKVDAFLFDENDDFFGIINVKASFAERRTDDVPMS
EALNNRGYYSALWTMDCKANPGPNPYNKGELGSPEGKRSAKRKDIEDEGYFTACFSYNQN
TQPTSRNRATKGQVIICDFKDPDDAFVEHIVEAYQEFMS

Figure 19: a) the CspAI RM system genes were found in convergent arrangement within the genomic DNA. Nucleotide and amino acid sequence of b) M.CspAI and c) R.CspAI.

As depicted in Figure 19a two genes coexist within *Corynebacterium* genome in convergent arrangement. Consequently, cloning of the DNA fragment containing both genes under a common promoter was not possible, and each gene was cloned to a separate vector.

4.2 Bioinformatic analysis of the CspAI RM system components

4.2.1 R.CspAI sequence analysis

Once the sequences of CspAI RM system were identified, bioinformatic analysis was performed. R.CspAI and R.AgeI are a pair of isoschizomers which recognize and cleave the same sequence,

5’ A/CCGGT 3’, but they exhibit different sensitivities to DNA methylation modification (Table 14).Crystal structure of R.AgeI was recently solved, thus a sequence alignment of R.CspAI and R.AgeI could provide a first clue of the R.CspAI cleavage mechanism. The active site of R.AgeI consists of residues E97, D142, K168 and D178. Alignment against R.CspAI has revealed identical residues, E224, D269, K286 and D296 (Figure 20).

Methyl-sensitive	**Methyl -insesitive**	**Recognition site**
AgeI	CspAI	5’ A/C**C**GGT 3’

Table 14: R.CspAI and R.AgeI are isoschizomers which recognize and cleave the site A/CCGGT (/ indicates the cleavage) but when the second C is methylated, R.AgeI is blocked while R.CspAI is active. In bold is depicted the site of methylation

All CCGG-family REs characterized so far share the residues from the conserved R-(D/E)R motif to recognize the central CCGG tetranucleotide. Surprisingly, CCGG recognition by AgeI differs from the enzymes belonging to the same family. Only E173 and R174 residues corresponding to the two last residues (underlined) from the R-(D/E)R motif are conserved in the AgeI sequence (Figure 21). The first arginine of the R-(D/E)R motif (underlined) which is used for the recognition of G in the first C:G bp in the other CCGG-family enzymes is replaced by K200 in AgeI (Tamulaitiene et al. 2016). So the catalytic core of AgeI is composed of E173, R174 and K200 for the recognition of C:G bp in the major groove and Q86 in minor groove. R.CspAI acquires identical residues of the catalytic core, E291, R292 and K318. For the recognition of A:T bp R.AgeI use residues K200, E214 and K224 in major groove and R90 in minor groove. Respective amino acids are found in R.CspAI sequence. In particular K318, E330 and K341 for major groove and R217 for minor groove. Crystal structure of R.AgeI has also revealed a unique element (residues 197-224) which appeared identical residues with R.CspAI (Figure 20). Overall, R.CspAI and R.AgeI share 22% identical amino acids.

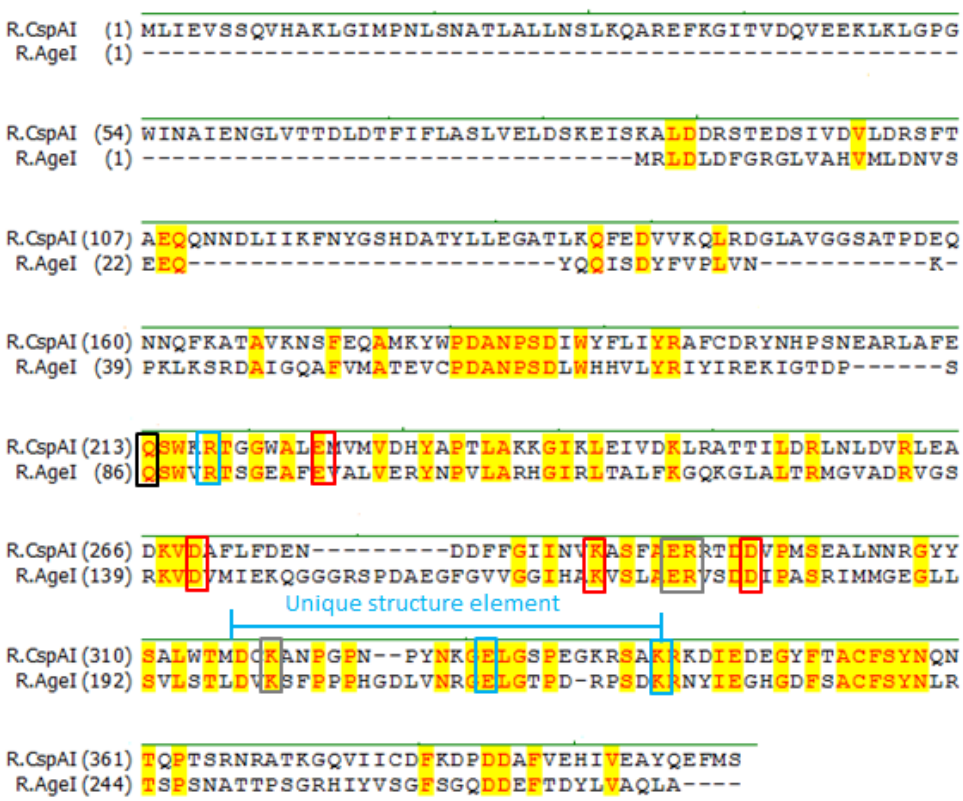


Figure 20: Sequence alignment of R.CspAI and R.AgeI. In red boxes are enclosed the residues of the active center (E224/E97, D269/D142, K286/K168 and D296/D178,R.CspAI/R.AgeI respectively). In gray boxes are the amino acids for the recognition of C:G bp in the major groove, E291/E173,R292/ R174 and K318/K200. The black box contains Q213/ Q86, responsible for the recognition of C:G in minor groove. In blue boxes are the residues for the recognition of outer A:T bp , E330/E214, K318/K200 and K341/K224, and R217/R90 for the minor groove. The unique structure element of R.AgeI is depicted by the blue line from 197 to 224.

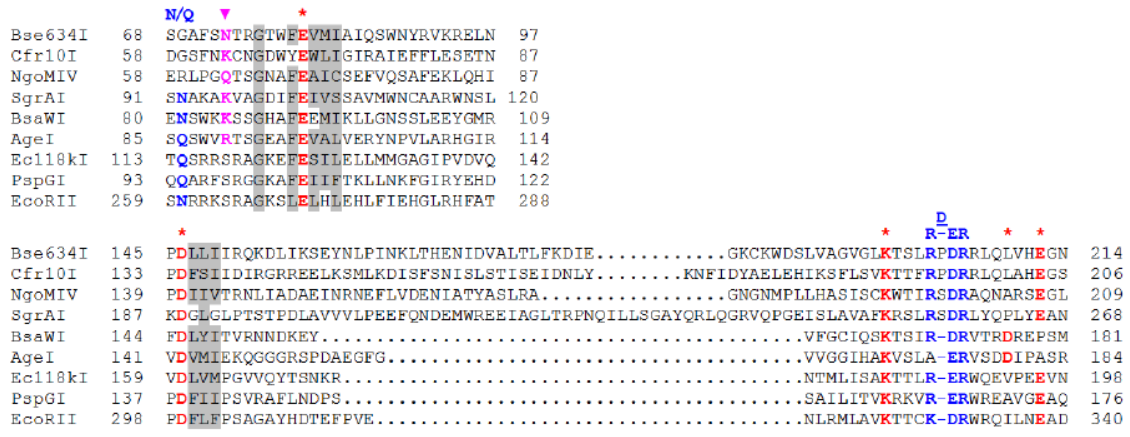


Figure 21: Conserved sequence motifs of the CCGG-family REs containing residues belonging to the active site (colored red marked by red stars), CCGG recognition (blue) and minor groove contact with the outer base pair (colored magenta and marked by magenta triangle) (Tamulaitiene et al. 2016).

Sequence alignment of R.CspAI was also performed with other CCGG-family enzymes (Figure 22). R.AgeI and R.BsaWI share identical amino acids in their active sites and in particular, D178 of R.AgeI is overlapped with D175 of R.BsaWI in the crystal structure. R.CspAI sequence contains these residues as well and this is a strong evidence of the conserved structure of R.CspAI active site.

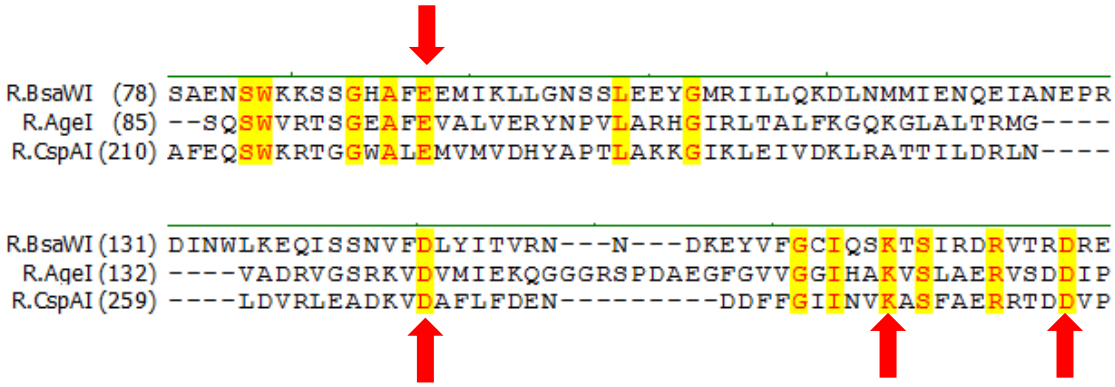


Figure 22: Sequence alignment of R.CspAI with R.AgeI and R.BsaWI. The yellow background with red letters depicts identical amino acids. The arrows show the conserved residues constituting the active center of R.AgeI and R.BsaWI.

4.2.2 M.CspAI sequence analysis

After M.CspAI amino acid sequence identification, a prediction of the methyltransferase type was performed using the Pfam software (17). The result showed that M.CspAI is a N4 or N6 Methyltransferase (Figure 23).

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites
				Start	End	Start	End	From	To				
N6_N4_Mtase	DNA methylase	Family	CL0063	29	277	30	275	2	229	231	119.2	2.1e-34	n/a

Figure 23: According to Pfam software prediction M.CspAI is a N4-N6 methyltransferase.

M.CspAI amino acid sequence alignment with other Mtases will show us the presence or absence of characteristic motifs. Motif IV is the most conserved motif among Mtases and it is considered to be their active center. Motif I is also conserved and it constitutes the AdoMet binding domain of the Mtases. Motif IV of N4-N6 Mtases is composed of PP(Y/F) while in C5 Mtases contains the PPC tripeptide. Motif I consists of FxGxG residues in all types of Mtases.

M.CspAI was aligned with representative proteins of each category of Mtases (Figure 24). M.HhaII is a N6-Mtase, M.AgeI and M.HpaII are C5 Mtase and M.SmaI is a N4-Mtase. Alignment showed that M.CspAI shares the highest homology with M.SmaI and it does has PP(Y/F) active site sequence characteristic in motif IV as most N4-N6 Mtases do (Figure 24).

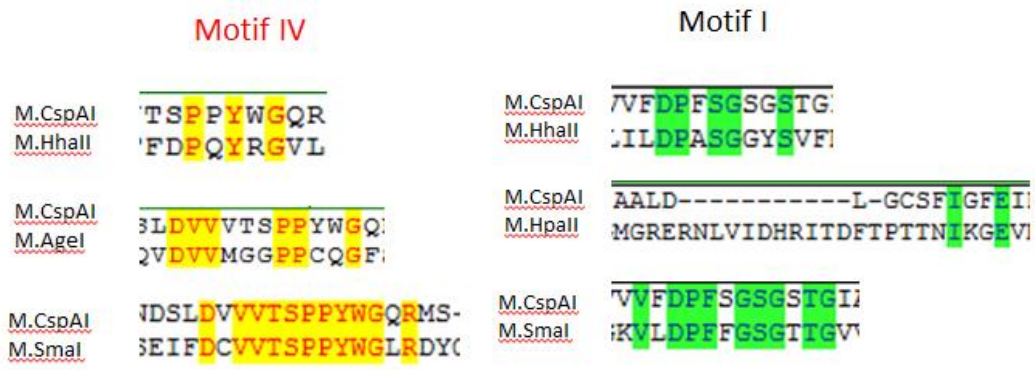


Figure 24: Sequence alignment for M.CspAI with M.HhaII (N6 Mtase), M.AgeI (C5 Mtase), M.HpaII (C5 Mtase) and M.SmaI (N4 Mtase) for Motif IV (PP(Y/F/C) and Motif I (FxGxG).

Apart from Motif IV and Motif I, M.CspAI shares Motif VI and Motif X with M.SmaI (Figure 25).

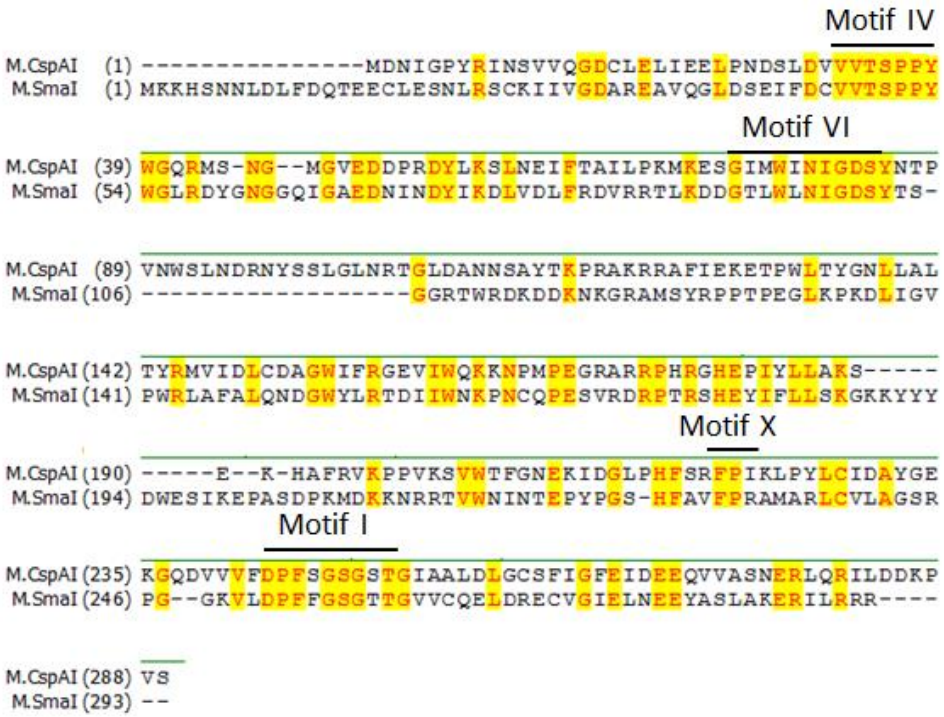


Figure 25: M.CspAI –M.SmaI sequence alignment revealed identical residues and conserved motifs between the two enzymes.

4.3 Cloning of the target enzymes

Restriction enzymes cloning and expression remains a challenge since they are toxic for host's genome. Their main function is to protect the host organism by recognizing specific sequence

and cutting the invading DNA. The host organism expresses a MT which recognizes the same sequence as the RE and methylates it so as to protect the genomic DNA from cleavage. Hence, respective Mt is appropriate for heterologous expression of a RE.

4.3.1 Cloning R.CspAI

R.CspAI gene was cloned in pET16b vector so as to carry a His tag at its N-terminus end. At first, traditional cloning method was followed (see Materials and methods) but it was proven unsuccessful. Therefore, the desired pET16b-R.CspAI plasmid construct was obtained by NEBuilder Assembly kit (Figure 26).

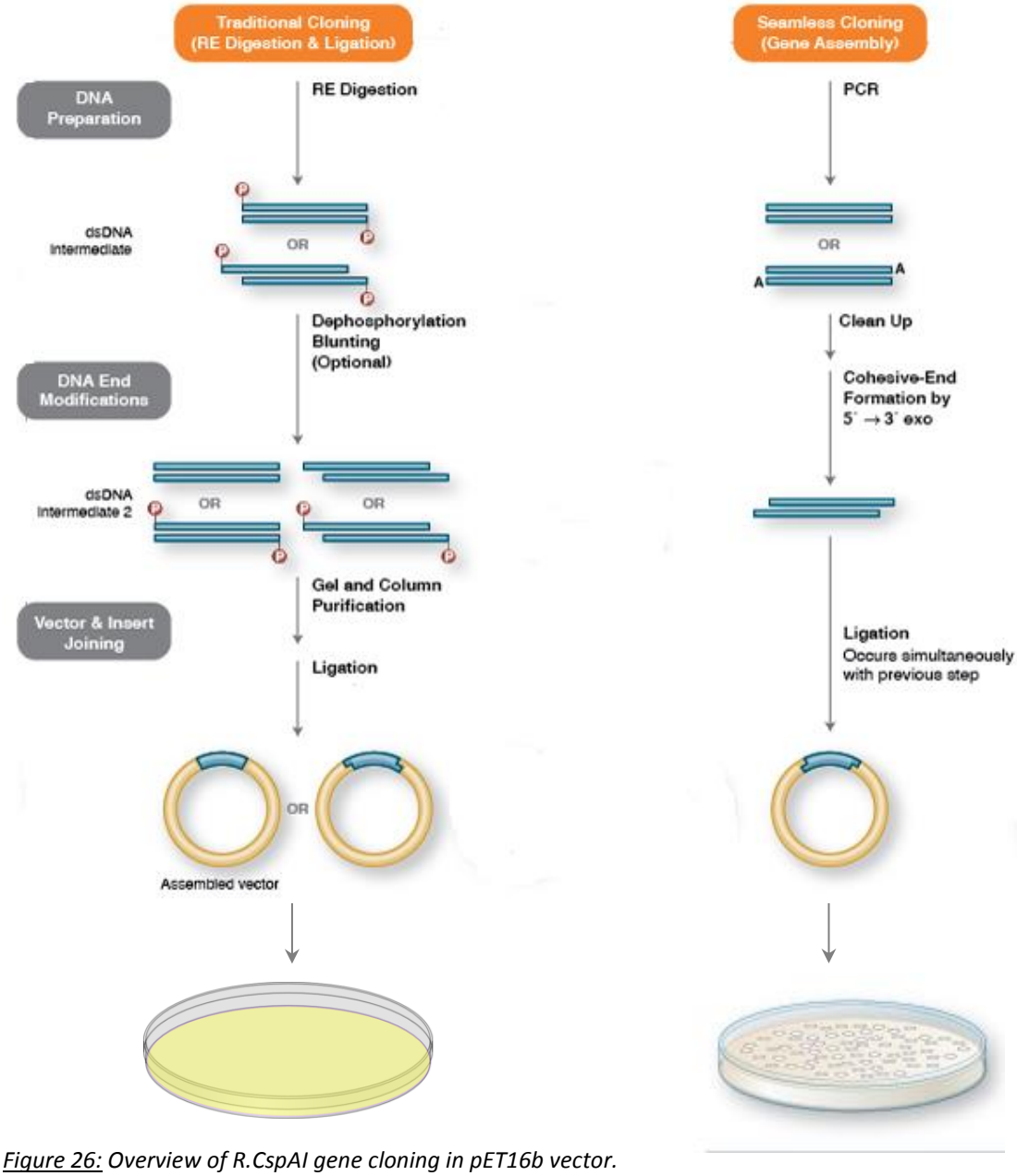
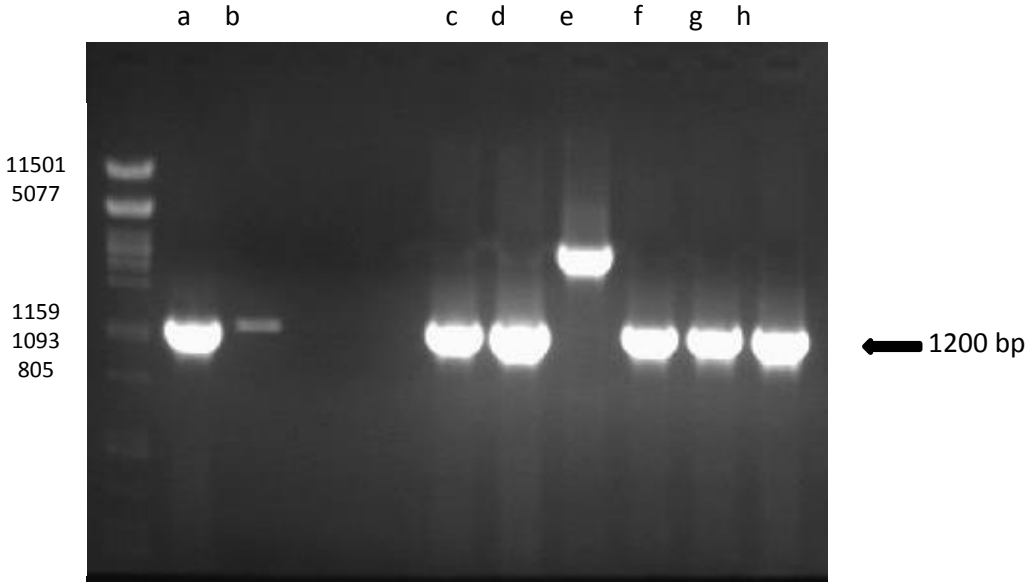


Figure 26: Overview of R.CspAI gene cloning in pET16b vector.

After the successful cloning the colonies were tested by colony PCR (Figure 27).

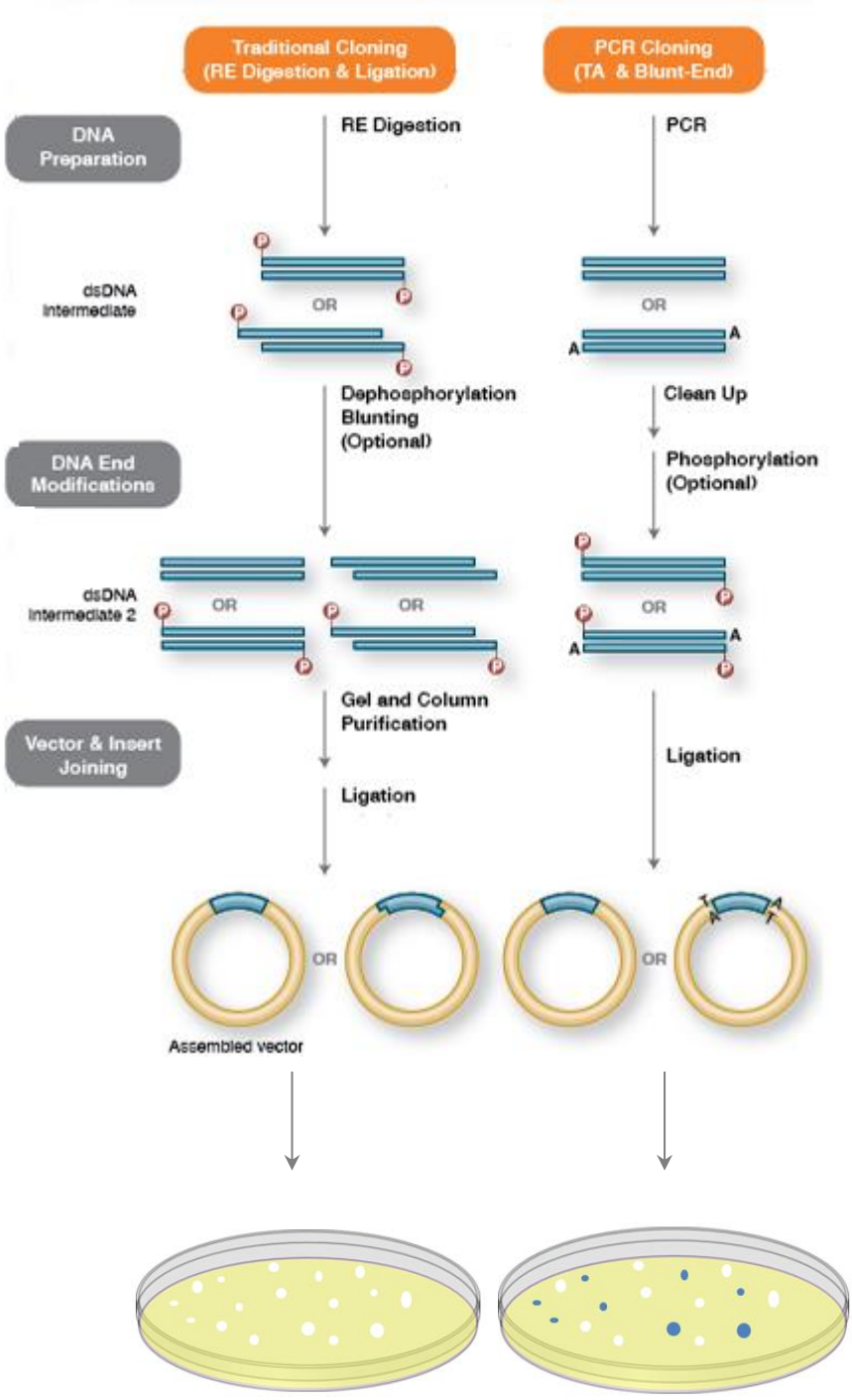


<u>Figure 27:</u> Gel agarose from colony PCR for testing the positives colonies containing the pET16b-R.CspAI plasmid construct. Lane a is the positive control using as template the genomic DNA of Corynebacterium and lane b is negative control containing the plasmid pET16b. The amplified colonies are lanes c-h.

Positives colonies c, d and f were selected and send for sequencing. Only colony f contained the correct pET16b-R.CspAI construct.

4.3.2 Cloning M.CspAI

The attempt for traditional cloning of M.CspAI in pACYC184 resulted in false positive colonies. Consequently, pGEM T-easy cloning kit was used for the construction of pACYC184-M.CspAI plasmid construct (Figure 28).



<u>Figure 28:</u> Workflow of cloning the M.CspAI gene. Cloning approaches are described in detail in materials and methods (sections 3.2.2 and 3.2.3).

Positive colonies of four variants of M.CspAI (fragments produced by primers F1, F2, F3 and F4) were verified by sequencing. Only F1 (containing only the gene M.CspAI) and F4 (designed based on RBS Library Calculator) contained the pACYC184-M.CspAI plasmid construct while F2

(including 43 nucleotides same as the 1875-1917 region of pLysE vector) and F3 (including RBS and NdeI site) colonies were false positive. As a result, the experiment was continued with these two types of M.CspAI. Unfortunately, sequencing results revealed mutations in the M.CspAI gene, thus site directed mutagenesis was performed to replace the mutated amino acids (Figure 29).

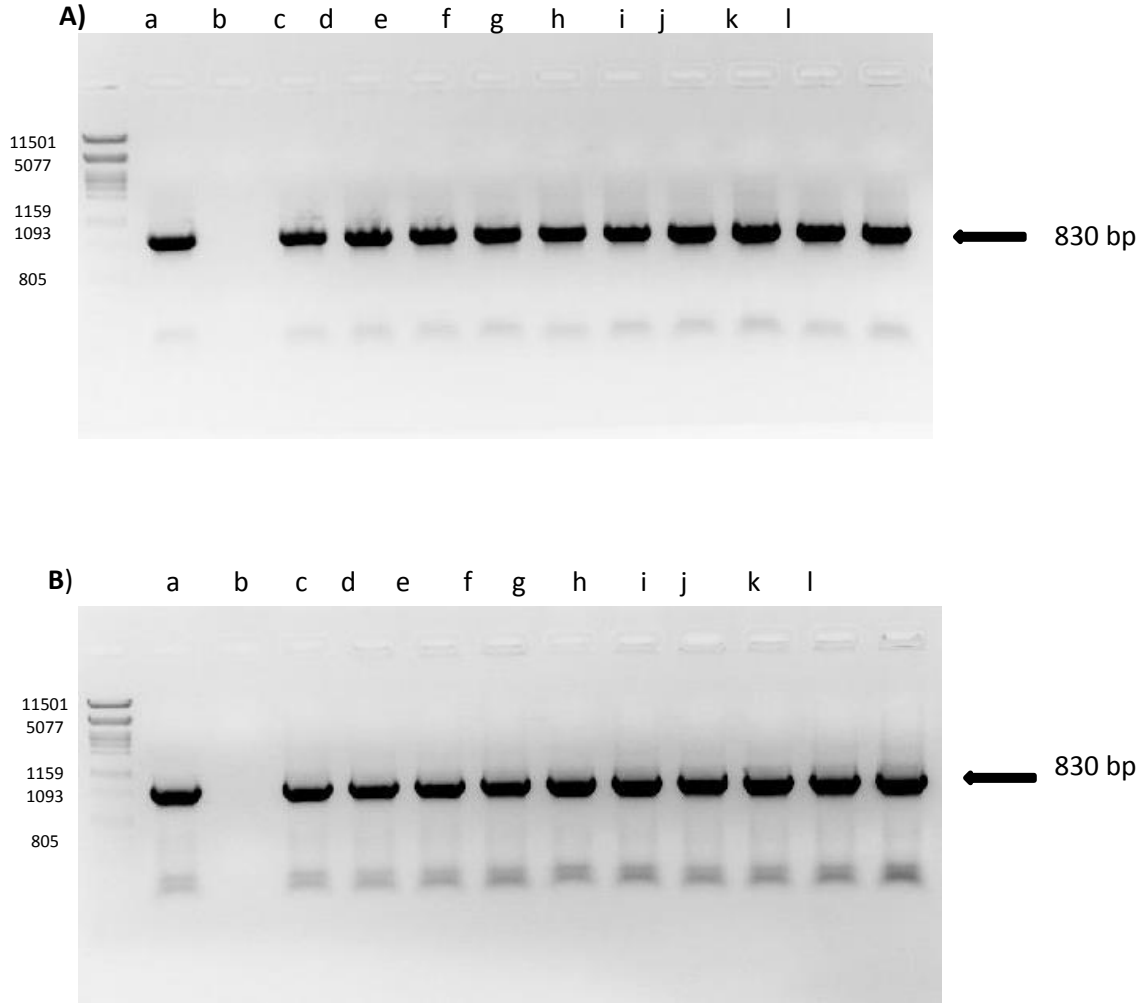


Figure29: Colony PCR for wt M.CspAI gene with A) F1 primers and B) F4 primers. A, B) Lanes a and b are positive and negative control,respectively, and lanes c-l are the selected colonies.

Two colonies of each wt M.CspAI variant were selected for sequencing and the results verified the wild type M.CspAI of each variant.

4.4 Heterologous expression of the R.CspAI protein

When T7 express cells were transformed simultaneously with the two plasmid constructs, pET16b-R.CspAI and pACYC184-wt M.CspAI, colonies didn't grow either because the transformation efficiency was reduced or due to the toxicity of the RE. Hence, competent T7 express cells supplemented with pACYC184-wt M.CspAI (F1 or F4) were transformed by pET16b-R.CspAI plasmid construct. Colonies did grow and were verified by PCR colony for R.CspAI gene (Figure 30).

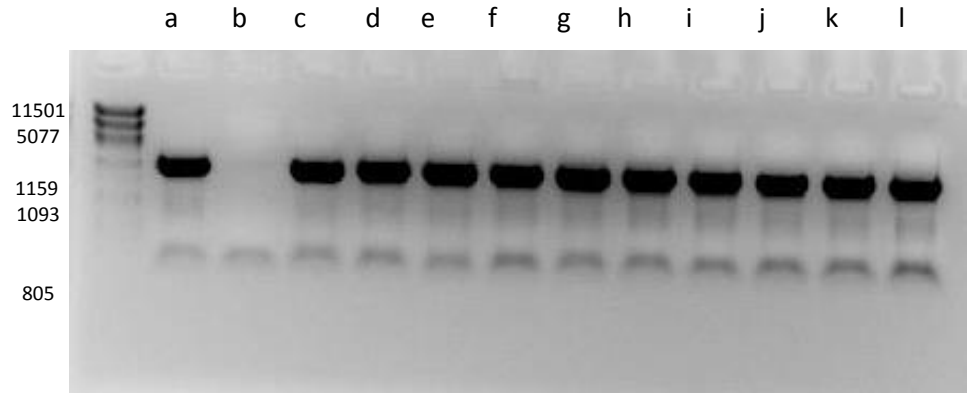


Figure 30: Colony PCR for positive colonies of the transformation in T7 express cells supplemented with pACYC184-M.CspAI. Lanes a and b are the positive and negative control respectively. Lanes c-g are colonies containing the pACYC184-wt F1 M.CspAI and pET16b-R.CspAI plasmid constructs while in lanes h-l are depicted the colonies with pACYC184-wt F4 M.CspAI and pET16b-R.CspAI constructs.

Three colonies of each M.CspAI were selected for small scale overexpression experiments.

Cell pastes were collected and lysed and subsequent samples were analyzed by Western Blot.

In Figure 31 sample a is T7 express cell without any construct after overnight incubation at 16^{o}C. Sample b and e are T7 express cells containing pACYC184-wt F1 M.CspAI and pACYC184-wt F4 M.CspAI, respectively. Sample c is T7 express cells with pACYC184-wt F1 M.CspAI and pET16b-R.CspAI 3h after induction and sample d is 16h after induction. Samples f and g are T7 express cells with pACYC184-wt F4 M.CspAI and pET16b- R.CspAI constructs 3h and 16h after induction, respectively. Samples h and i are positive controls, containing His-RFP and His-GFP accordingly. The Western blot confirms the expression of R.CspAI (samples c,d,f,g) with both variants of wt M.CspAI (F1, F4) at the induction time tested.

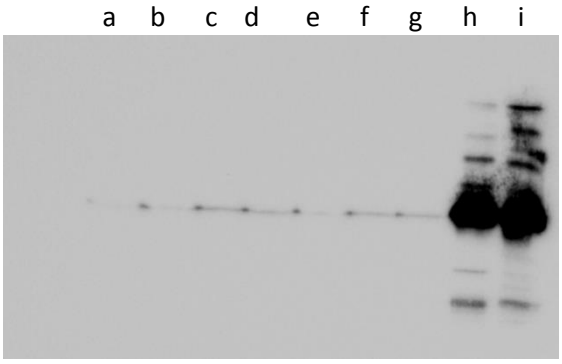
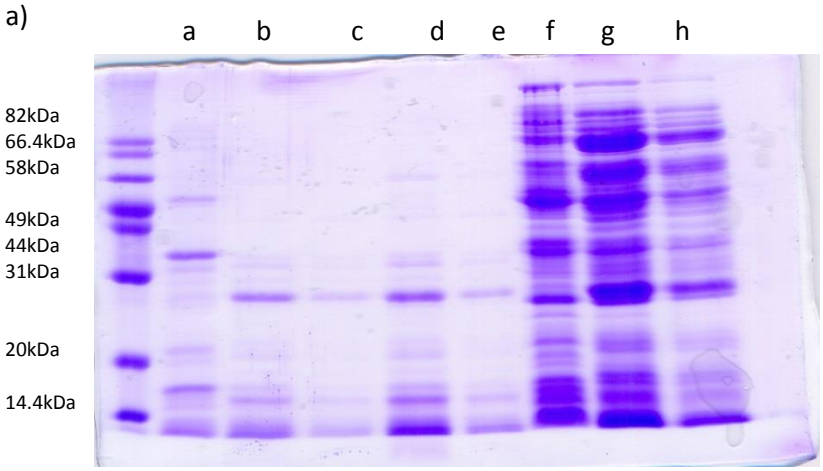
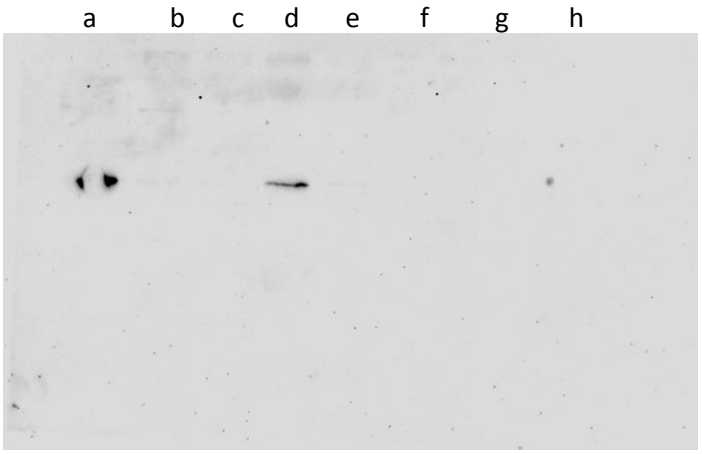


Figure 31: Western blot with anti-His antibody for His-R.CspAI identification.

A large scale overexpression of R.CspAI was performed and the samples were lysed, purified by Ni-NTA column and analyzed by 12.5% SDS-PAGE and Western blot. Figure 32 depicts the SDS-PAGE and Western blot analyzing the same samples. Sample a is the load of T7 express cells with pACYC184-wt F1 M.CspAI and pET16b-R.CspAI. Samples b and c are elution 1 and 2 from T7 express cells containing pACYC184-wt F1 M.CspAI and pET16b-R.CspAI, respectively. Samples d and e are elution 1 and 2 from T7 express cells containing pACYC184-wt F4 M.CspAI and pET16b-R.CspAI. Sample f is T7 express cells with pACYC184-wt F1 M.CspAI and samples g and h are elution 1 and 2 from the sample f. As it is indicated from figure 32a R.CspAI was not identified by SDS-PAGE analysis. Although, Western blot exhibit a signal in sample d where the R.CspAI coexisted with wt F4 M.CspAI.



b)



<u>Figure 32:</u> a) 12.5% SDS-PAGE and b) Western blot after induction and purification with Ni-NTA column. The sample order was the same in both analysis processes.

5. Discussion-Perspectives

Isoschizomers are REs which recognize and cleave the same DNA sequence but they origin from different bacterial strains. Furthermore there are pairs of isoschizomers which exhibit different sensitivities to DNA methylation modifications. A pair of this type is CspAI and AgeI which recognize and cleave sequence 5'-A/CCGGT-3'. When the first C is methylated both enzymes are blocked but when the second C is methylated, AgeI is blocked while CspAI is active. CspAI RM system DNA and protein sequences were unknown until now while the crystal structure of R.AgeI has been recently discovered.

The identification of CspAI RM system enabled cloning of the RE and the respective MT to selected *E.coli* strain. The heterologous expression of R. CspAI was achieved in T7 express *E.coli* cells and was verified by western blot. Further optimization of the expression scheme will be tested (time point of induction, incubation time, temperature, broth composition) in order to proceed with the purification of the target protein. Due to the presence of a His-tag, Ni-NTA affinity chromatography will be used along with other types of chromatography such as ion exchange and/or gel filtration. The final protein sample should be free of other DNA modifying activities (non-specific exo-, endonucleases).

Of great importance will be as well to elucidate the methylation sensitivity of R.CspAI. According to Rebase (http://rebase.neb.com) its activity has been examined only for M.HpaII and M.MspI DNA methylation. So, it should also be tested for M.EcoBI and M.EcoKI DNA methylation so as to provide a complete characterization of the enzyme.

Sequence alignment of R.CspAI and R.AgeI that were performed in the current thesis revealed identical residues in their active sites. Taking account the unique cleavage mechanism that R.AgeI has adapted it would be of great importance to proceed with the elucidation and comparison of R.CspAI crystal structure. In addition, the knowledge of the three dimensional structure of the enzyme will provide crucial information about the active site and the critical residues responsible for the methylation sensitivity.

The RM system is composed of the RE and the respective Mtase. Comparative analysis of M.CspAI sequence has provided the first evidence for classification of M.CspAI as an N4-N6 MT based on conserved motifs among MTs. Expression and purification of pure protein samples will be the goals of future studies in order to fully characterize M.CspAI.

Numerous pairs of isoschizomers have been identified but only the crystal structure of one RE of each pair has been solved. Consequently, crystallization of R.CspAI and structural comparison of the isoschizomers R.CspAI –R.AgeI will be the first of each kind. In addition, R.CspAI crystal structure will reveal the key residues for methylation sensitivity. These data will set the basis for rational engineering approaches in order to alter methylation sensitivities of selected DNA modifying enzymes.

Bibliography

1. Bheemanaik, Shivakumara, Yeturu V. R. Reddy, and Desirazu N. Rao. 2006. “Structure, Function and Mechanism of Exocyclic DNA Methyltransferases.” *Biochemical Journal* 399 (2): 177–90.
2. BioLabs, New England. 2012. “Key RestRiction Enzymes foR DnA MethylAtion AnAlysis,” 1–3.
3. Bujnicki, Janusz M. 2002. “Sequence Permutations in the Molecular Evolution of DNA Methyltransferases.” *BMC Evolutionary Biology* 2: 1–11.
4. Khulan, Batbayar, Reid F. Thompson, Kenny Ye, Melissa J. Fazzari, Masako Suzuki, Edyta Stasiek, Maria E. Figueroa, et al. 2006. “Comparative Isoschizomer Profiling of Cytosine Methylation: The HELP Assay.” *Genome Research* 16 (8): 1046–55.
5. Kong, Shiao-tong, Foreign Application, and Priority Data. 2011. “(12) United States Patent” 2 (12): 12–15.
6. Kosinski, Jan, Marcin Feder, and Janusz M. Bujnicki. 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.
7. Kumar, Sanjay, Xiaodong Cheng, Saulius Klimasauskas, Sha Mi, Janos Posfai, Richard J Roberts, and Geoffrey G Wilson. 1994. “The DNA (Cytosine-5) Methyltransferases” 22 (1): 1–10.
8. Loenen, Wil A.M., David T.F. Dryden, Elisabeth A. Raleigh, Geoffrey G. Wilson, and Noreen E. Murrayy. 2014. “Highlights of the DNA Cutters: A Short History of the Restriction Enzymes.” *Nucleic Acids Research* 42 (1): 3–19.
9. Malone, Thomas, Robert M. Blumenthal, and Xiaodong Cheng. 1995. “Structure-Guided Analysis Reveals Nine Sequence Motifs Conserved among DNA Amino-Methyltransferases, and Suggests a Catalytic Mechanism for These Enzymes.” *Journal of Molecular Biology* 253 (4): 618–32.
10. NEBuilder HiFi DNA Assembly Master Mix / NEBuilder HiFi DNA Assembly Cloning Kit
11. Pingoud, A., M. Fuxreiter, V. Pingoud, and W. Wende. 2005. “Type II Restriction Endonucleases: Structure and Mechanism.” *Cellular and Molecular Life Sciences* 62 (6): 685–707.
12. Pingoud, Alfred, and Albert Jeltsch. 2001. “AND SUMMARY SURVEY AND Structure and Function of Type II Restriction Endonucleases” 29 (18): 3705–27.
13. Pingoud, Alfred, Geoffrey G. Wilson, and Wolfgang Wende. 2014. “Type II Restriction Endonucleases - A Historical Perspective and More.” *Nucleic Acids Research* 42 (12): 7489–7527.
14. Tamulaitiene, Giedre, Virginija Jovaisaite, Gintautas Tamulaitis, Inga Songailiene, Elena Manakova, Mindaugas Zaremba, Saulius Grazulis, Shuang Yong Xu, and Virginijus Siksnys. 2016. “Restriction Endonuclease AgeI Is a Monomer Which Dimerizes to Cleave DNA.” *Nucleic Acids Research* 45 (6): 3547–58.

15. Timinskas, A, V Butkus, and A Janulaitis. 1995. “Sequence Motifs Characteristic for DNA and DNA Methyltransferases. Classification of All DNA Methyltransferases.” *Gene* 157: 3–11.
16. Vasu, K., and V. Nagaraja. 2013. “Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense.” *Microbiology and Molecular Biology Reviews* 77 (1): 53–72.
17. https://pfam.xfam.org/

Supplementary data

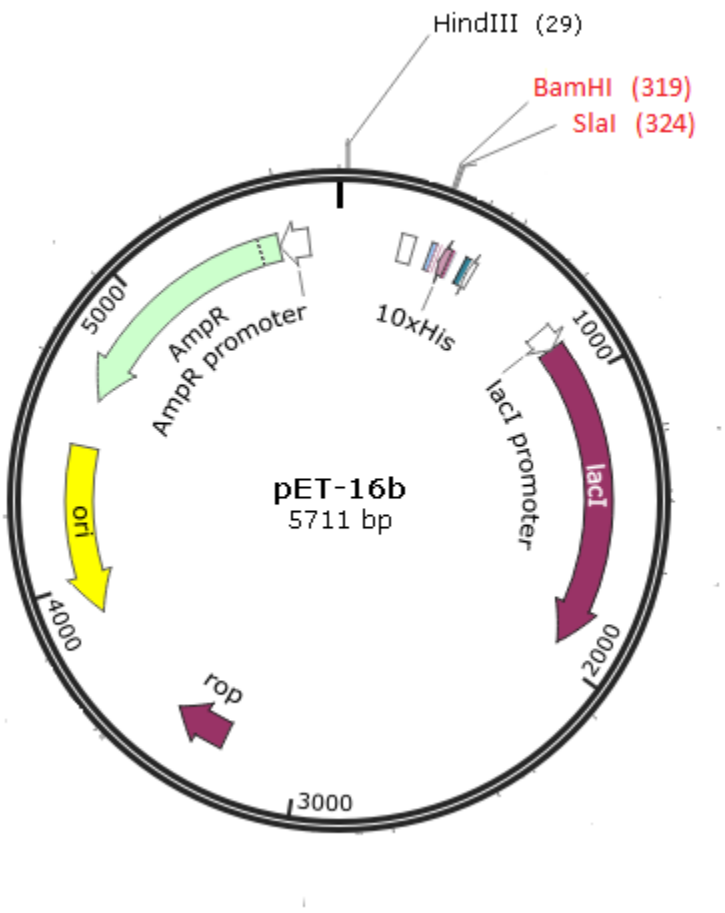


Figure 1: pACYC184 plasmid map. In red are depicted the sites of M.CspAI cloning.

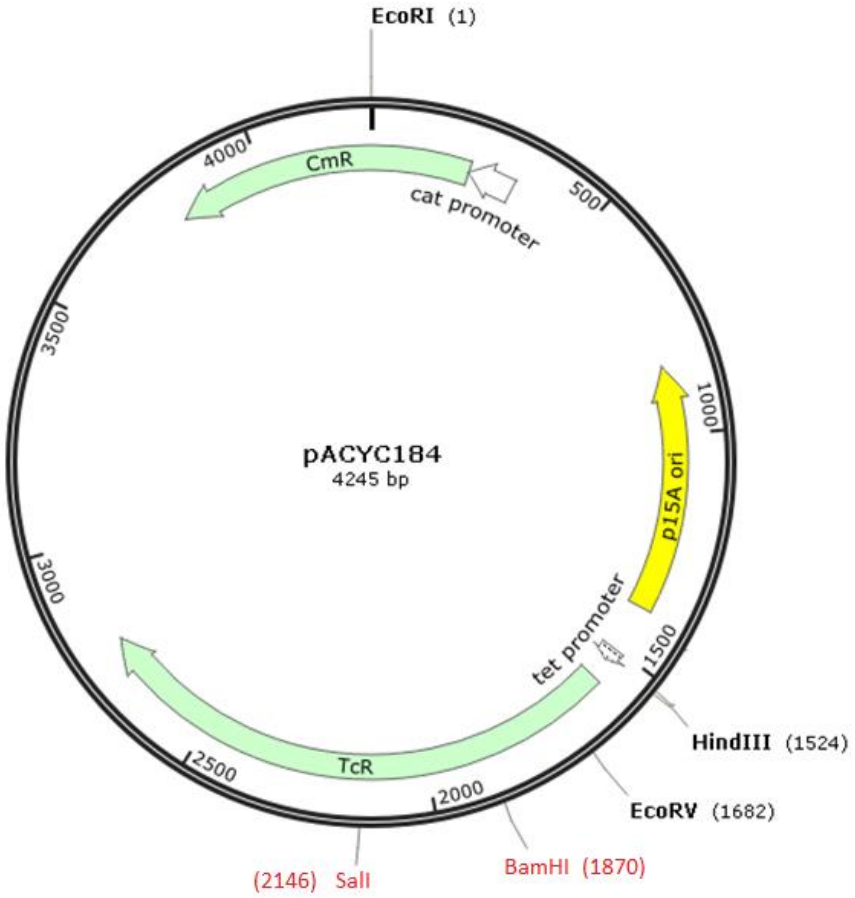


Figure 2: pET16b plasmid map. In red are depicted the cloning sites of R.CspAI.

Dh5a	F^- endA1 glnV44 thi-1 recA1 relA1 gyrA96 deoR nupG purB20 ϕ80dlacZΔM15 Δ(lacZYA-argF)U169, **hsdR17($r_K^-m_K^+$),** λ^-
RR1	F- D(gpt-proA)62 leuB6 supE44 ara-14 galK2 lacY1 **D(mcrC-mrr)**rpsL20 xyl-5 mtl-1 recA13 Strr RecA+
HB101	F- D(gpt-proA)62 leuB6 supE44 ara-14 galK2 lacY1 **D(mcrC-mrr**)rpsL20(StrR) xyl-5 mtl-1 recA13
NEB 5-alpha $F'I^q$	F´ proA+ B+ lacIq Δ(lacZ)M15 zzf::Tn10 (TetR) / fhuA2Δ(argF-lacZ) U169 phoA glnV44 f80D(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 **hsdR17**
T7 express	fhuA2 lacZ::T7 gene1 [lon] ompT gal sulA11 R(mcr-73::miniTn10--TetS)2 [dcm] R(zgb-210::Tn10--TetS) endA1 Δ(mcrC-mrr)114::IS10

Table 1: Genotypes of cells used in this thesis. In red are depicted the genes which are responsible for DNA methylation in specific sites and should be considered before for cloning a RE in bacterial cells.