GLORIA ALEJANDRA GALLO ENAMORADO

CARACTERIZAÇÃO ESTRUTURAL DA PROTEÍNA CSM2, UMA PROTEÍNA DO SISTEMA CRISPR-CAS

Tese apresentada à Universidade Federal de São Paulo, para obtenção do título de Doutor em Ciências.

São José dos Campos
2018

GLORIA ALEJANDRA GALLO ENAMORADO

CARACTERIZAÇÃO ESTRUTURAL DA PROTEÍNA CSM2, UMA PROTEÍNA DO SISTEMA CRISPR-CAS

Tese apresentada à Universidade Federal de São Paulo, para obtenção do título de Doutor em Ciências.

Orientador:
Prof. Dr. Marcelo Alves da Silva Mori
Co-orientador:
Prof. Dr. Martin Rodrigo Alejandro Würtele Alfonso

São José dos Campos
2018

Gallo Enamorado, Gloria Alejandra

Caracterização estrutural da proteína Csm2, uma proteína do sistema CRISPR-Cas. / Gloria Alejandra Gallo Enamorado. -- São José dos Campos, 2018.

Tese (Doutorado) - Universidade Federal de São Paulo. Instituto de Ciência e Tecnologia. Programa de Pós-Graduação Interunidades em Biotecnologia.

Título em inglês: Structural characterization of Csm2, a protein related to CRISPR-Cas system.

1. Expressão de proteínas recombinantes. 2. Cristalização de proteínas. 3. Difração de raios X. 4.CRISPR-Cas. 5. RNAi. 6. Csm2. 7. *Thermotoga maritima*

UNIVERSIDADE FEDERAL DE SÃO PAULO
INSTITUTO DE CIÊNCIA E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

Chefe do Departamento:
Prof. Dr. Eduardo Antonelli
Coordenadora do curso de pós-graduação:
Profa. Dra. Cláudia Barbosa Ladeira de Campos

GLORIA ALEJANDRA GALLO ENAMORADO

CARACTERIZAÇÃO ESTRUTURAL DA PROTEÍNA CSM2, UMA PROTEÍNA DO SISTEMA CRISPR-CAS

Presidente da banca:
Prof. Dr. Marcelo Alves da Silva Mori

Banca examinadora:
Prof. Dr. João Bosco Pesqueiro
Prof. Dr. Katlin Brauer Massirer
Prof. Dr. Jörg Kobarg
Prof. Dr. Andrey Fabrício Ziem Nascimento

Gostaria dedicar este trabalho a meu pai Jorge, minha mãe Irma Gloria e minhas irmãs Daniela e Natalia, que mesmo na distância, encontram-se presentes no meu dia a dia.

Gostaria de agradecer aos Professores Marcelo Mori, Martin Würtele, Cláudia Campos e André Zelanis, pela orientação e pelas oportunidades oferecidas, a todos os professores, colegas e técnicos do Programa de Pós-Graduação em Biotecnologia pela ajuda, troca de conhecimentos e motivação e aos responsáveis pelas linhas de difração de raios-X MX-1 e MX-2 do Laboratório Nacional de Luz Síncrotron pelo apoio técnico e científico.

Este trabalho foi financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (projeto FAPESP 11/50963-4), o Conselho Nacional de Desenvolvimento Científico e Tecnológico, (projetos CNPq 480411/2011-5 e 448833/2014-0) e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) financiando uma bolsa de doutorado.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS, SIGLAS, DEFINIÇÕES

Å	Ångström, 10^{-10} m
Cas	CRISPR *associated proteins* (proteínas associadas à CRISPR)
Cascade	CRISPR-*associated complex for antiviral defence* (complexo de defesa antiviral associado à CRISPR)
CRISPR	*Clustered Regularly Interspaced Short Palindromic Repeats* (repetições palindrômicas curtas agrupadas e regularmente interespaçadas)
crRNA	CRISPR RNA
crRNP	CRISPR *ribonucleoprotein (*ribonucleoproteína CRISPR*)*
Da	Dalton
DNA	*Deoxyribonucleic acid* (ácido desoxirribonucleotídeo)
dNTP	*Deoxyribonucleotide triphosphate* (desoxirribonucleotídeos fosfatados)
DTT	ditiotreitol
FPLC	*Fast Protein Liquid Chromatography* (cromatografia líquida rápida de proteína)
HEPN	*Higher Eukaryotes and Prokaryotes Nucleotide*
in silico	através de uma simulação computacional
in vitro	processos biológicos que têm lugar fora dos sistemas vivos
IPTG	Isopropil-β-D-1-tiogalactopiranosídeo
kDa	Kilo Dalton
L	litro
LB	meio de cultura Luria Bertani
LC-MS	*Liquid Chromatography Mass Spectrometry* (espectrometria de massa de cromatografia líquida)
locus	local fixo em um cromossomo (plural *loci*)
MAD	*Multiple Anoumalous Diffraction (difração anômala com comprimentos de onda múltiplos)*
mAU	*milli Arbitrary Units (*unidades mili-arbitrárias*)*
MCS	*Multiple Cloning Site (*sítio de múltipla clonagem*)*
MIR	*Multiple Isomorphous Replacement (substituição isomórfica múltipla)*
M	mol/L
min	minuto
MR	*Molecular Replacement (*substituição molecular*)*

mRNA	RNA mensageiro
NCBI	*National Center for Biotechnology Information*
NIH	*National Institute of Health*
NMR	*Nuclear Magnetic Resonance (ressonância magnética nuclear)*
nt	Nucleotídeo
OD	*Optic density* (densidade ótica)
Óperon	conjunto de genes dos procariontes, contíguos e controlados coordenadamente.
PAM	*Protospacer Adjacent Motif* (motivo adjacente do proto-espaçador)
pb	pares de bases
PCR	*Polymerase Chain Reaction* (reação em cadeia da polimerase)
PDB	*Protein Data Bank*
pH	Potencial de hidrogênio
PI	PAM *interacting*
PMSF	*phenylmethanesulfonylfluoride* (fluoreto de fenilmetano sulfonil)
precrRNA	precursor de CRISPR RNA
q.s.p.	quantidade suficiente para
RAMP	*Repeat-Associated Mysterious Protein* (proteína misteriosa associada a repetições)
RMSD	*Root-Mean-Square Deviation*
RNA	*Ribonucleic acid* (ácido ribonucleotídico)
RNAi	RNA de interferência
RNP	Ribonucleoproteína
r.p.m.	rotações por minuto
RRM	RNA *Recognition Motif (*motivo de reconhecimento de RNA*)*
SAD	*Single-wavelength Anomalous Dispersion* (difração anômala simples)
SAXS	*Small-Angle X-ray Scattering* (espalhamento de raios X a baixos ângulos)
SDS PAGE	*Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis* (eletroforese em gel de poliacrilamida com sulfato dodecil de sódio)
SIR	*Single Isomorphous Replacement (substituição isomórfica simples)*
TEV	*Tobacco Etch Virus (*Vírus da gravura do tabaco*)*
Tm	*Melting temperature (ponto de fusão médio)*
tracrRNA	crRNA de ativação em *trans*

Tris	tris-(hidroximetil)-aminometano
U	Unidade de atividade enzimática
Vm	coeficiente de Matthews

ABREVIATURAS DOS RESÍDUOS DE AMINOÁCIDOS

A	Ala	Alanina
C	Cys	Cisteína
D	Asp	Aspartato
E	Glu	Glutamato
F	Phe	Fenilalanina
G	Gly	Glicina
H	His	Histidina
I	Ile	Isoleucina
K	Lys	Lisina
L	Leu	Leucina
M	Met	Metionina
N	Asn	Asparagina
P	Pro	Prolina
Q	Gln	Glutamina
R	Arg	Arginina
S	Ser	Serina
T	Thr	Treonina
V	Val	Valina
W	Trp	Triptofano
Y	Tyr	Tirosina

RESUMO

O sistema CRISPR-Cas é constituído pelas sequências CRISPR (*Clustered Regularly Intespaced Short Palindromic Repeats*) junto com as proteínas Cas (CRISPR *associated*) e tem como função conferir às bactérias e archaeas proteção contra DNA e RNA exógenos através de um mecanismo baseado em RNA de interferência (RNAi). Este sistema consiste em um óperon composto por repetições (*repeats*) de sequências idênticas entre si, separadas de sequências variáveis (*spacers*) provenientes dos ácidos nucleicos invasores, junto com proteínas Cas. O objetivo principal deste projeto foi clonar, expressar e purificar de forma recombinante a proteína Csm2 de *Thermotoga maritima* MSB8, relacionada ao sistema CRISPR-Cas, visando sua caracterização bioquímica e estrutural. A proteína Csm2 integra um complexo de ribonucleoproteínas (RNP) Cas menos estudadas, denominado complexo Csm, que está envolvido no processamento de DNA e RNA exógeno. A proteína foi produzida de forma recombinante em *Escherichia coli* BL21(DE3) e purificada por cromatografia de afinidade e filtração em gel. Após concentração, a proteína cristalizou no grupo espacial $P3_121$ numa célula unitária com as dimensões a=77 Å b=77 Å c=160 Å α=90° β=90° γ=120°. A estrutura de Csm2 foi solucionada via difração anômala simples de cádmio a uma resolução de 2,4 Å e comprimento de onda de 1,458 Å. A estrutura revela que a Csm2 está composta de uma α-hélice longa de 42 aminoácidos rodeada de três α-hélices menores. Esta estrutura mostra que a proteína é capaz de formar dímeros devido a sua extensa superfície de contato atribuída por sua longa α-hélice. Esta interação é adicionalmente estabilizada pela hélice N-terminal, que é inserida dentro da porção helicoidal C-terminal da subunidade adjacente. A dimerização de Csm2 foi auxiliarmente confirmada por cromatografia de exclusão molecular da proteína pura recombinante, seguido de análises por espectrometria de massas das frações eluídas. Devido a seu papel no complexo ribonucleoproteíco CRISPR do tipo Csm, a estrutura da proteína Csm2 é importante para o entendimento do mecanismo de ação do sistema CRISPR-Cas subtipo III-A dentro do contexto dos diferentes sistemas CRISPR-Cas.

ABSTRACT

The clusters of regularly interspaced short palindromic repeats (CRISPR) and the Cas (CRISPR-associated) proteins form an adaptive immune system in bacteria and archaea that evolved as an RNA-guided interference mechanism to target and degrade foreign genetic elements. This system consists of an operon composed of regions of identical repeats, separated by variable spacers derived from invading nucleic acids. Together with the Cas proteins, this system forms an adaptive and heritable immune system. The main objective of this project was to clone, recombinantly express and purify the Csm2 protein from *Thermotoga maritima* MSB8, for biochemical and structural characterization. This protein is part of a marginally studied complex of Cas ribonucleoproteins (RNP) termed the Csm complex, which is involved in the targeting of exogenous DNA. The protein was recombinantly produced in *Escherichia coli* BL21(DE3) and purified by affinity and gel filtration chromatography. After concentration, the protein crystallized in space group $P3_121$ with an unit cell with dimensions a=77 Å b=77 Å c=160 Å α=90° β=90° γ=120°. Csm2 was solved via cadmium single wavelength anomalous diffraction phasing at 2,4 Å resolution at a wavelength of 1,458 Å. The structure reveals that Csm2 is composed of a large 42 amino-acid long α-helix flanked by three shorter α-helices. The structure also shows that the protein is capable of forming dimers mainly via an extensive contact surface conferred by its long α-helix. This interaction is further stabilized by the N-terminal helix, which is inserted into the C-terminal helical portion of the adjacent subunit. The dimerization of Csm2 was additionally confirmed by size exclusion chromatography of the pure recombinant protein followed by mass spectrometry analysis of the eluted fractions. Because of its role in the Csm CRISPR RNP complex, the crystal structure of Csm2 is of great importance for clarifying the mechanism of action of the subtype IIIA CRISPR-Cas system, in the context of the different CRISPR-Cas systems.

1. INTRODUÇÃO

1.1 O SISTEMA CRISPR-CAS: FUNÇÃO

1.1.1 Descoberta e definições

O sistema CRISPR-Cas de bactérias e archaeas é constituído pelas sequencias CRISPR (do inglês *Clustered Regularly Interspaced Short Palindromic Repeats,* repetições palindrômicas curtas agrupadas e regularmente interespaçadas*)* e pelas proteínas Cas *(CRISPR associated,* associadas às CRISPR*).* Ele tem como função conferir proteção contra DNA e RNA exógenos através de um mecanismo baseado em RNA de interferência (RNAi).

O locus CRISPR foi inicialmente descrito por Francisco Mojica *et al.* (Mojica *et al.*, 1993, Mojica *et al.*, 1995, Mojica *et al.*, 2000, Mojica *et al.*, 2005). Os CRISPR consistem em curtas sequências repetitivas de DNA intercaladas por sequências espaçadoras derivadas de bacteriófagos ou de plasmídeos conjugados (figura 1). Este fato levou à hipótese surpreendente que as sequências CRISPR poderiam estar relacionadas à defesa contra bacteriófagos (Mojica *et al.*, 2005). Outros grupos de pesquisa chegaram a conclusões semelhantes na mesma época (Bolotin *et al.*, 2005, Pourcel *et al.*, 2005, Barrangou *et al.*, 2007).

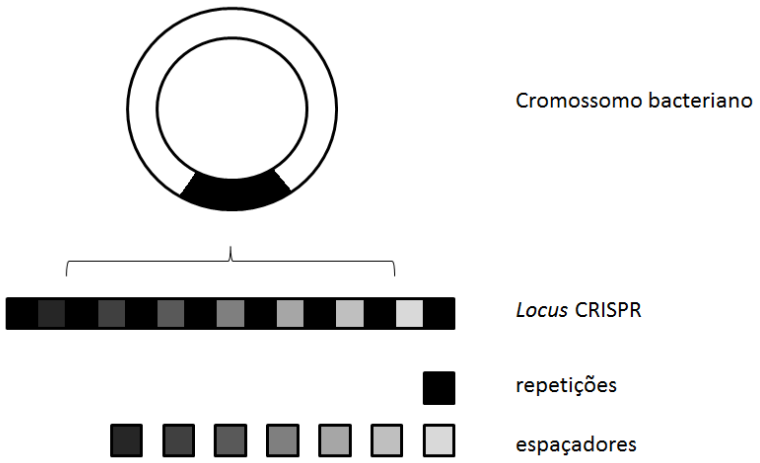


Figura 1. CRISPR.

Fragmentos do DNA invasor (espaçadores) incorporados no arranjo de repetições do locus CRISPR do cromossomo bacteriano.

O sistema CRISPR-Cas consiste em repetições de fragmentos de 21 a 37 pb interespaçados por sequências não repetitivas com tamanho de 25 a 57 pb, sendo que muitas espécies comportam dois ou mais *loci* CRISPR que por sua vez geralmente se encontram próximos ou adjacentes aos genes das proteínas Cas,

indicando que estes genes estão funcionalmente relacionados (Jansen *et al.*, 2002, Brouns *et al.*, 2008). A relação das sequencias CRISPR com os genes Cas está esquematizada na figura 2.

Figura 2. Relação CRISPR-Cas.

O locus CRISPR consiste em repetições palindrômicas de 21-37pb (quadrados em preto), separados de sequências espaçadoras de 25-57pb (quadrados em tons cinza) dentro do genoma bacteriano. Os espaçadores não possuem traços característicos nas suas sequências. Os genes Cas (setas cinza) se encontram próximos aos loci CRISPR e adjacentes a uma sequência líder (seta branca).

1.1.2 Mecanismos de ação

Após análise computacional, um esquema hipotético de mecanismo de ação foi proposto para o complexo efetor CRISPR-Cas. O sistema CRISPR-Cas possui três estágios característicos, conhecidos como aquisição, expressão e interferência. Seguindo este modelo, os procariotos adquirem e integram fragmentos de genes provenientes de fagos dentro do arranjo dos espaçadores das sequências CRISPR. Numa etapa subsequente estes fragmentos, junto com as repetições, são expressas para formar um RNA de interferência (atualmente conhecido como crRNA ou CRISPR RNA) que junto com proteínas Cas forma um complexo que tem como objetivo degradar DNA alvo, gerando assim um mecanismo de imunidade hereditária (Makarova *et al.*, 2006). Este mecanismo está esquematizado na figura 3.

A demonstração experimental da imunidade adaptativa foi inicialmente relatada por Barrangou *et al.* (Barrangou *et al.*, 2007) em experimentos envolvendo a proteção de fagos em *S. thermophilus*. Neste trabalho foi comprovada a integração de um fragmento de DNA de um fago (este fragmento de DNA invasor foi nomeado pouco tempo depois como proto-espaçador) dentro do arranjo CRISPR permitindo a proteção em um ataque posterior de esse mesmo fago (figura 4). Isto permitiu ao mesmo grupo de pesquisa introduzir uma primeira aplicação no campo da biotecnologia: Imunização de bactérias utilizadas na indústria de laticínios contra infecção por fagos (Barrangou and Horvath, 2012).

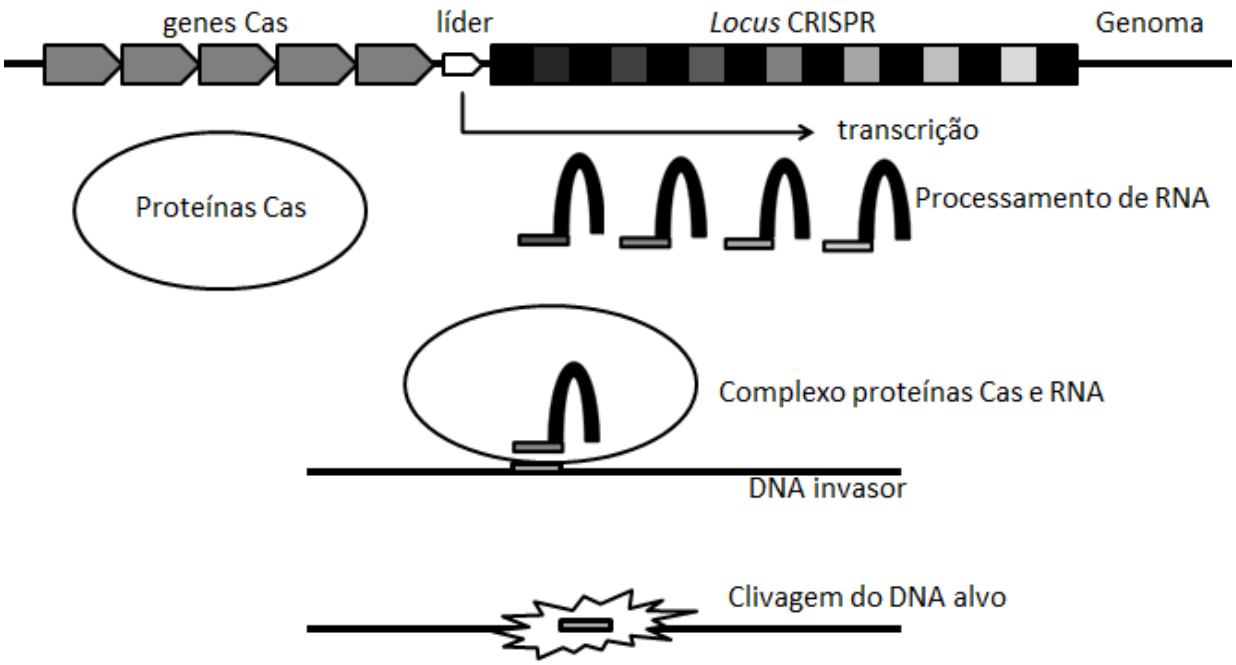


Figura 3. Modelo básico do sistema CRISPR-Cas.

O locus CRISPR é transcrito por meio de uma sequência líder que contém o promotor do óperon, logo após acontece o processamento do RNA. Este RNA forma um complexo com as proteínas Cas. O complexo contendo o RNA derivado do DNA invasor se liga de forma complementar à sequência do DNA alvo (fago ou plasmídeo conjugado) provocando a sua degradação.

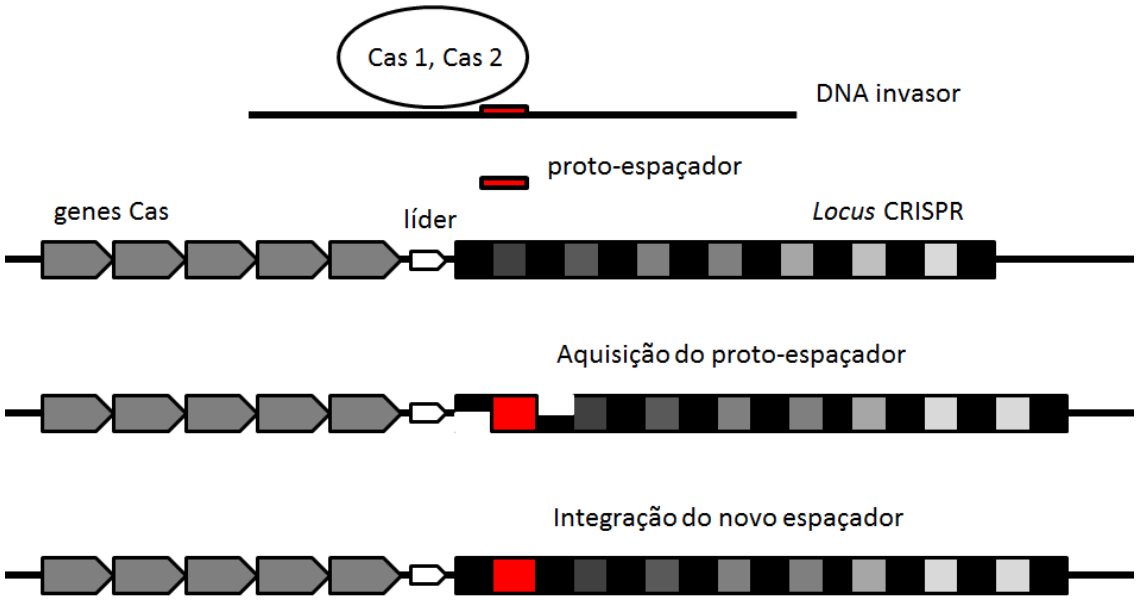


Figura 4. Primeiro estágio da imunidade adaptativa do CRISPR: A aquisição.

Este passo começa pelo reconhecimento do DNA invasor pelas proteínas Cas1 e Cas2. Um fragmento de DNA invasor é clivado formando o proto-espaçador. Logo este proto-espaçador é inserido no início do locus CRISPR formando um novo espaçador.

O mecanismo pertinente de expressão do sistema CRISPR-Cas foi inicialmente detalhado por Brouns *et al.* (Brouns *et al.*, 2008). Foi possível confirmar que as sequências espaçadoras do sistema CRISPR-Cas de uma cepa *Escherichia coli* K12 são transcritas como pequenos RNAs e são utilizadas por proteínas Cas da bactéria para mediar uma resposta antiviral que neutraliza a infecção por fagos. Após a transcrição do óperon CRISPR, proteínas do assim denominado complexo Cascade (do organismo *Escherichia coli* K12), especificamente a proteína Cas6 cliva um precursor de crRNA (pre-crRNA) criando um crRNA amadurecido com aproximadamente 57 nucleotídeos. O complexo Cascade liga o crRNA para formar um complexo crRNP (ribonucleoproteína CRISPR) (figura 5) e recruta a proteína Cas3 para degradar o ácido nucleico invasor usando o crRNA como guia (Brouns *et al.*, 2008, Richter *et al.*, 2012) (figura 6).

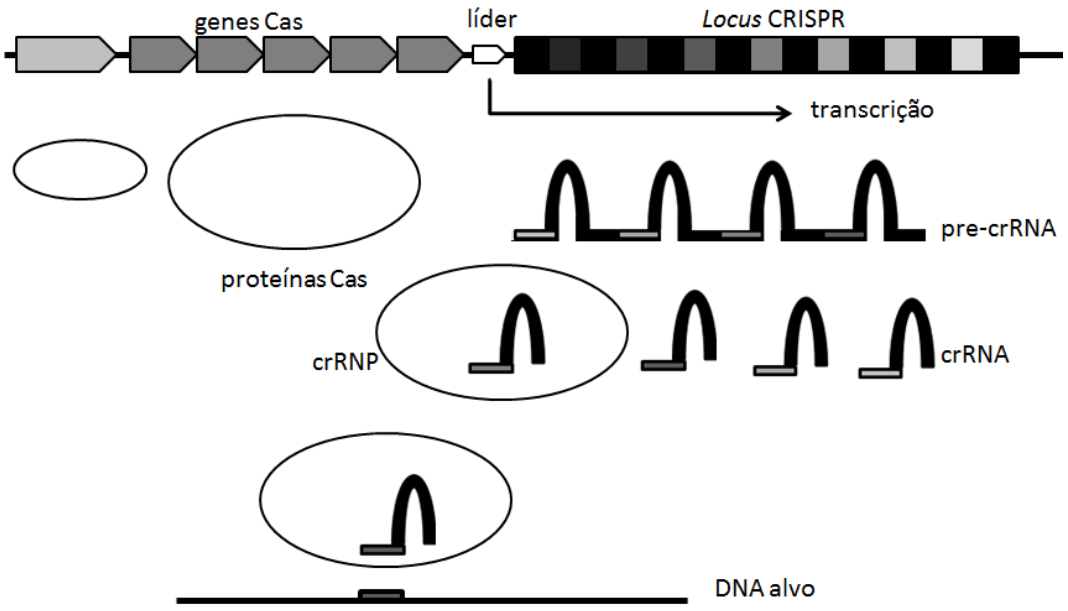


Figura 5. Segundo estágio da imunidade adaptativa do CRISPR: A expressão.

O locus CRISPR é transcrito e o pre-crRNA é processado em pequenos crRNA por proteínas Cas. O crRNA maduro e as proteínas Cas se agrupam para formar um complexo crRNP (ribonucleoproteína CRISPR). Este crRNA constitue um guia para futuro reconhecimento de DNA invasor. Na bactéria *Escherichia coli* K12, o Cascade é evidenciado pela agrupação de Cas8e, Cas11, Cas7, Cas5e e Cas6e (representados nos polígonos cinza escuro) formando um complexo com o crRNA.

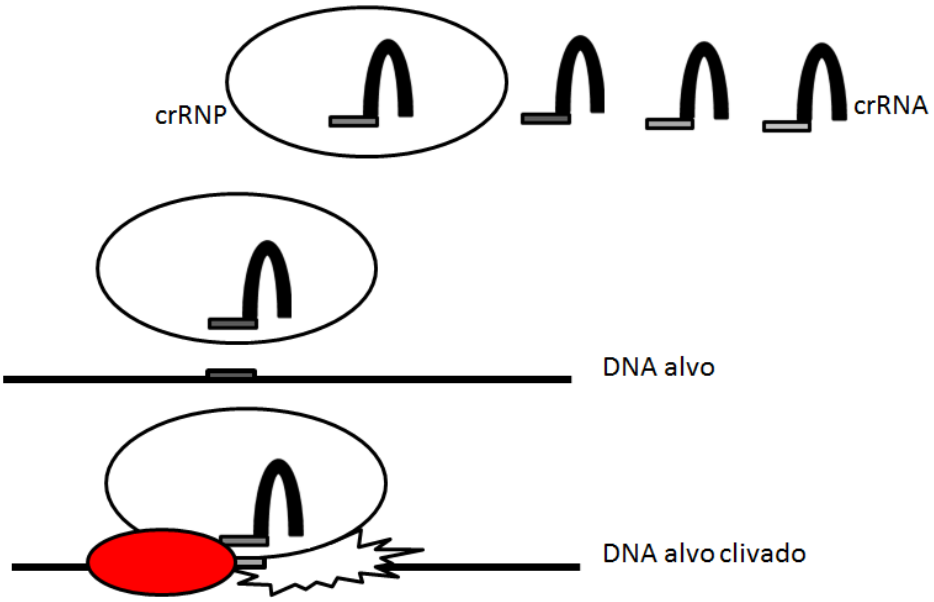


Figura 6. Terceiro estágio da imunidade adaptativa do CRISPR: A interferência.

O complexo crRNP contendo o crRNA derivado do DNA invasor se liga de forma complementária a sequência de DNA invasor e posteriormente o degrada. Em *Escherichia coli* K12, o crRNP (Cascade e crRNA) se liga a sequência do DNA invasor, a proteína responsável pela clivagem de este DNA corresponde a Cas3 (representada em vermelho), uma proteína com domínios com atividade nuclease e helicase.

Inicialmente, resultados indicavam que DNA é o principal alvo molecular da interferência por CRISPR-Cas (Marraffini and Sontheimer, 2008, Makarova *et al.*, 2011b). Entretanto, posteriormente foi constatado que alguns sistemas CRISPR são capazes de visar RNA como alvo (Hale *et al.*, 2009, Koonin *et al.*, 2017).

1.1.3 Classificação dos sistemas CRISPR-Cas

Diversos estudos têm sido realizados para sistematizar a diversidade dos sistemas CRISPR-Cas. Assim, em análises sobre a evolução e classificação dos mais diversos sistemas CRISPR-Cas foram unificadas diversas famílias de proteínas Cas, observando homologias entre elas, seja por comparação de sequências ou de estruturas (Makarova *et al.*, 2011a, Koonin *et al.*, 2017). Em todo caso, aparentemente todos os mecanismos do sistema CRISPR-Cas apresentam vias semelhantes de aquisição, expressão e interferência (Makarova *et al.*, 2011b).

Foi esboçada inicialmente uma classificação com uma única classe de proteínas e dividido em quatro tipos: I, II, III-A e III-B (Makarova *et al.*, 2011a, Makarova *et al.*, 2011b). Posteriormente a existência de novos sistemas CRISPR-

Cas levou à classificação dos diversos sistemas em duas classes: Sistemas da classe 1 (figura 7), formam um complexo RNP efetor composto de diversas subunidades. Estes sistemas são classificados em três tipos (I,III e IV) e pelo menos doze subtipos. Sistemas da classe 2, possuem uma única proteína efetora e são classificados em três tipos (II, V e VI). A figura 8 descreve os subtipos mais relevantes desta classe (Makarova *et al.*, 2015, Koonin *et al.*, 2017).

Os sistemas CRISPR-Cas da classe 1 e do tipo I possuem o gene Cas3 característico destes sistemas. Este gene codifica uma proteína que degrada DNA fita simples estimulada por uma helicase com capacidade para desenovelar tanto DNA dupla fita como híbridos de RNA e DNA (Sinkunas *et al.*, 2011, Gong *et al.*, 2014, Huo *et al.*, 2014). Usualmente o domínio helicase está fusionado a um domínio endonuclease envolvido na clivagem do DNA alvo (Sinkunas *et al.*, 2011, Gong *et al.*, 2014, Huo *et al.*, 2014, Mulepati and Bailey, 2011). O domínio HD encontra-se localizado no extremo amino terminal das proteínas Cas3.

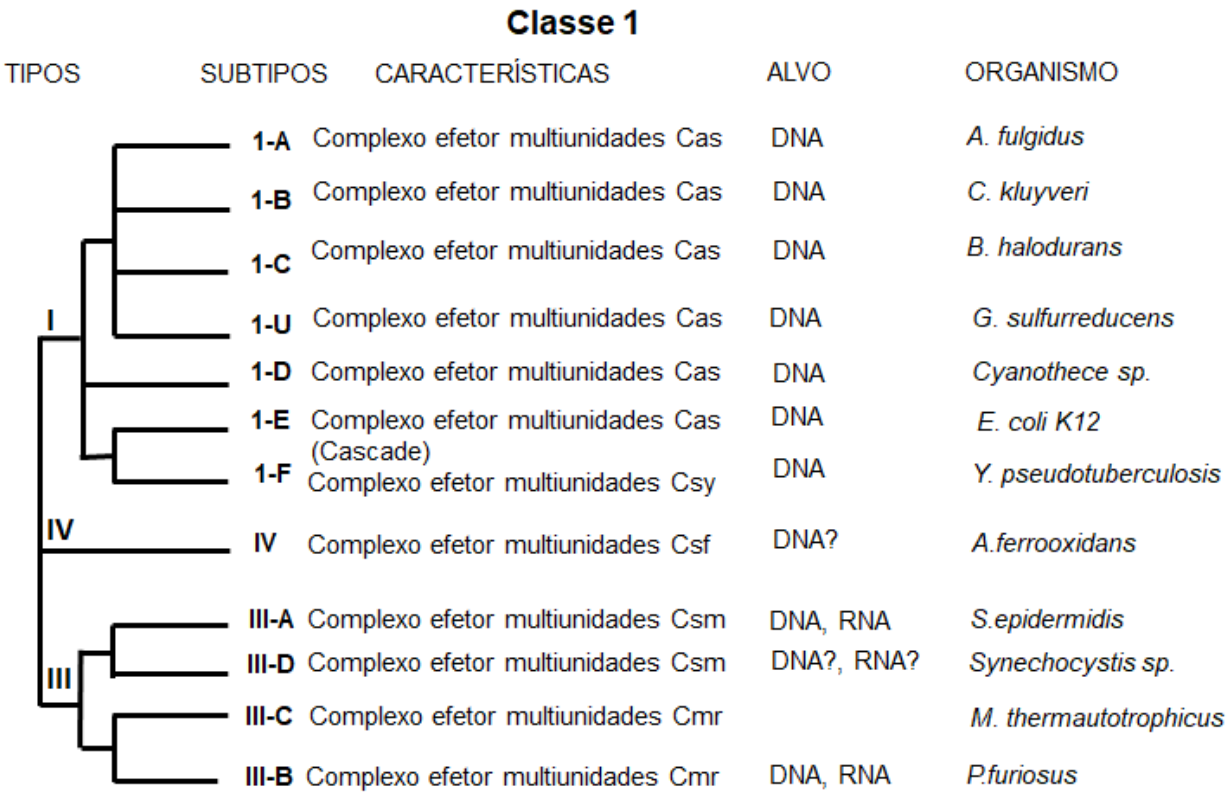


Figura 7. Classificação do sistema CRISPR classe 1.

O sistema classe 1 aborda o tipo mais comum e diverso (tipo I), o tipo III representado em várias archaeas porém menos frequente em bactérias e o tipo IV. Estudos citados: subtipo I-E (Brouns *et al.*, 2008), subtipo III-A (Marraffini and Sontheimer, 2008), subtipo III-B (Hale *et al.*, 2009), IV (Vestergaard *et al.*, 2014), outros subtipos (Makarova *et al.*, 2015, Koonin *et al.*, 2017).

Continuando na classe 1, o sistema tipo III foi classificado em quatro subtipos: III-A, III-B, III-C e III-D. O subtipo III-A foi anteriormente chamado de Mtube ou módulo Csm. O subtipo III-B, foi chamado de módulo Cmr ou módulo RAMP (*Repeat-Associated Mysterious Protein*). Estes subtipos podem ser facilmente distinguidos pela presença de diferentes genes específicos que codificam suas pequenas subunidades. Trata-se de Csm2 no caso particular do subtipo III-A e Cmr5 no caso do subtipo III-B. Inicialmente o subtipo III-A e o subtipo III-B eram conhecidos por clivar DNA (Marraffini and Sontheimer, 2008) e RNA (Hale *et al.*, 2009) respectivamente. Atualmente sabemos que ambos subtipos são polivalentes, clivando tanto DNA como RNA (Samai *et al.*, 2015, Hale *et al.*, 2012, Spilman *et al.*, 2013, Staals *et al.*, 2014, Tamulaitis *et al.*, 2014, Goldberg *et al.*, 2014, Deng *et al.*, 2013, Peng *et al.*, 2015). Mais recentemente foram descobertos dois novos subtipos para o sistema III, o III-C de módulos Cmr e o subtipo III-D de módulos Csm. (Koonin *et al.*, 2017).

O tipo IV desta classe ainda não foi caracterizado. Este sistema codifica um complexo efetor que consiste das proteínas Csf1 (similar a Cas8), Cas5 e uma única cópia da proteína Cas7.

A classe 2 inclui o sistema CRISPR-Cas tipo II, que difere dos tipos anteriores (figura 8). A proteína característica deste sistema é a endonuclease Cas9, uma proteína multidomínio que combina as funções de complexo crRNA-efetor com a clivagem do DNA alvo por ação dos domínios nucleases RuvC e HNH (Jinek *et al.*, 2012, Garneau *et al.*, 2010). Esta proteína também contribui no processo de expressão de nucleotídeos auxiliada por uma RNAse III para catalizar o processo de amadurecimento de pre-crRNA (figura 9B) (Heler *et al.*, 2015, Wei *et al.*, 2015). A maior parte dos *loci* tipo II codificam o tracrRNA (*trans activating* CRISPR RNA), que é parcialmente complementar às repetições do seu respectivo arranjo CRISPR e é importante para a maturação do crRNA e formação do complexo Cas9/crRNA (Jiang and Doudna, 2017, Deltcheva *et al.*, 2011). A formação de um dúplex entre o tracrRNA e o crRNA, guia a endonuclease Cas9 até seu alvo (figura 9B) (Chylinski *et al.*, 2014, Chylinski *et al.*, 2013, Briner *et al.*, 2014, Deltcheva *et al.*, 2011). O DNA alvo precisa conter um motivo PAM (*Protospacer Adjacent Motif)* que consiste em uma sequência de três nucleotídeos 5′-NGG-3′, onde “N” pode ser qualquer base nitrogenada, seguido de duas guaninas (G) (Anders *et al.*, 2014). Esta sequência

PAM é reconhecida pelo domínio PI (*PAM-interacting domain*) localizado no final do C-terminal da proteína Cas9 (Nishimasu *et al.*, 2014)

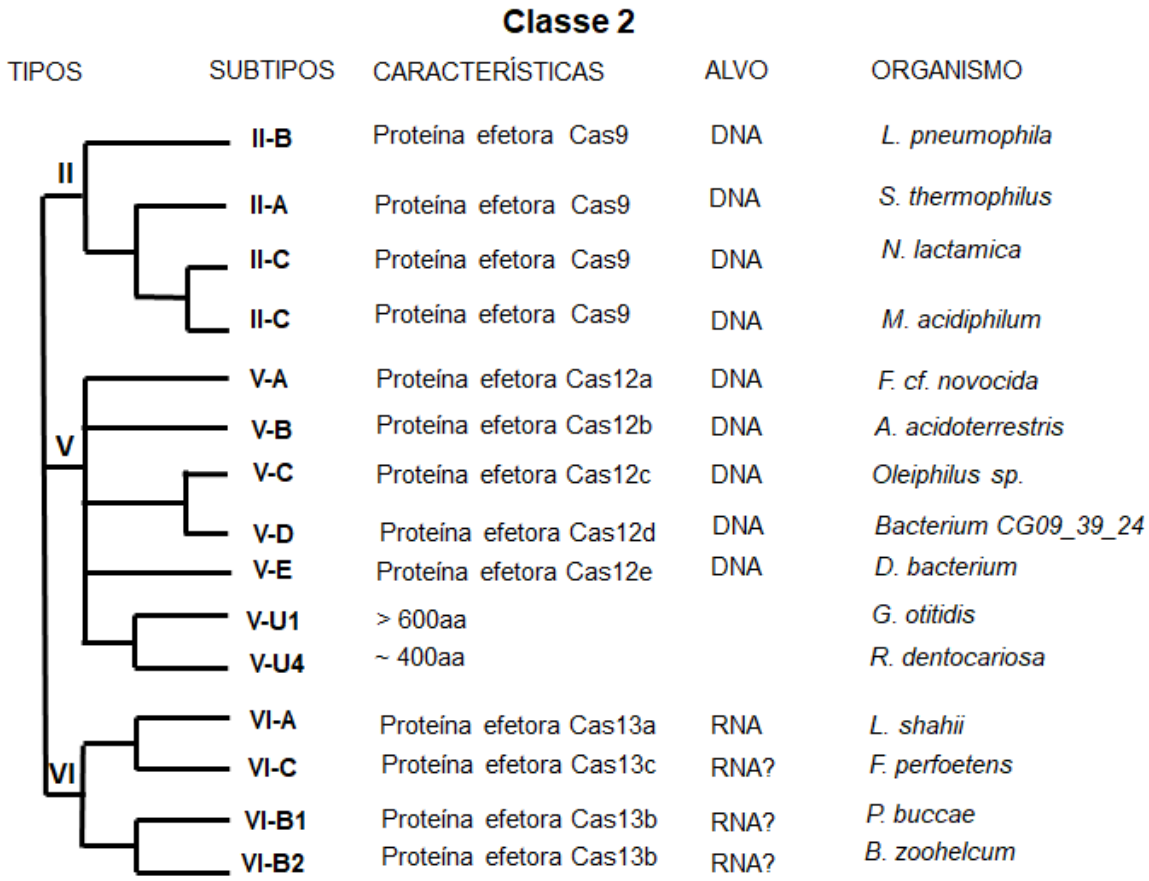


Figura 8. Classificação do sistema CRISPR classe 2.

O sistema classe 2 inclui o tipo II com sua proteína efetora Cas9 usada amplamente como ferramenta de edição de genomas, o tipo V com sua proteína efetora prevista Cas12 e recentemente o tipo VI ainda não caracterizado. Estudos citados: tipo II (Bolotin *et al.*, 2005, Barrangou *et al.*, 2007, Sapranauskas *et al.*, 2011, Gasiunas *et al.*, 2012, Deltcheva *et al.*, 2011, Jinek *et al.*, 2012, Cong *et al.*, 2013, Mali *et al.*, 2013), tipo V e VI (Zetsche *et al.*, 2015, Makarova *et al.*, 2015, Koonin *et al.*, 2017).

Em um estudo inicial, foi possível transferir o óperon CRISPR de uma espécie de bactéria para outra, gerando novas possibilidades de aplicações para este mecanismo do sistema tipo II (Sapranauskas *et al.*, 2011). Este resultado permitiu que outros estudos demonstrassem que proteínas Cas9 purificadas são capazes de clivar DNA alvo *in vitro* quando guiadas por crRNAs e tracrRNA (Jinek *et al.*, 2012, Gasiunas *et al.*, 2012). Por sua vez, isto abriu a estratégia de utilização do sistema CRISPR-Cas como uma ferramenta de engenharia genética, tornando possível

editar com sucesso o genoma de células de mamíferos, assim como de inúmeros outros sistemas (Sander and Joung, 2014, Kim *et al.*, 2014, Cong *et al.*, 2013, Cress *et al.*, 2015, Zheng *et al.*, 2016, Bohaciakova *et al.*, 2017, Dong *et al.*, 2018) e abrindo espaço para inúmeras aplicações.

Finalmente, os últimos tipos a serem descritos correspondem a sistemas CRISPR-Cas tipo V e tipo VI. A proteína característica do tipo V é a Cas12 (inicialmente descrita como Cf1) que contém um domínio nuclease semelhante a RuvC e homólogo ao domínio respectivo em Cas9, que cliva DNA. O tipo VI codifica o efetor Cas13, que integra dois domínios de ligação HEPN *(Higher Eukaryotes and Prokaryotes Nucleotide)* para o processamento de pre-crRNA. A superfamília HEPN consiste em RNAses que estão envolvidas em várias funções relacionadas à defesa de procariotos e eucariotos (Shmakov *et al.*, 2015). A descoberta de um efetor putativo contendo domínios HEPN levou à previsão de que o sistema CRISPR Cas tipo VI clive RNA (Shmakov *et al.*, 2015, Abudayyeh *et al.*, 2016). Devido à presença estimada de uma única subunidade no complexo crRNA-efetor, ambos sistemas, tipo V e tipo VI foram incluídos dentro da classe 2 de CRISPR-Cas (Makarova and Koonin, 2015, Chylinski *et al.*, 2014)

Em síntese, tanto o sistema CRISPR-Cas tipo I, como o sistema CRISPR-Cas tipo II e tipo V clivam DNA (Garneau *et al.*, 2010, Sinkunas *et al.*, 2011, Gasiunas *et al.*, 2012, Zetsche *et al.*, 2015). Entretanto, o sistema CRISPR-Cas tipo III pode clivar tanto DNA como RNA (Marraffini and Sontheimer, 2008, Hale *et al.*, 2009, Staals *et al.*, 2013, Staals *et al.*, 2014, Samai *et al.*, 2015, Taylor *et al.*, 2015). O mais recente sistema CRISPR-Cas descrito do tipo VI cliva RNA (Abudayyeh *et al.*, 2016).

Durante as fases de aquisição e interferência, os sistemas CRISPR-Cas tipo I e II precisam da participação de uma sequência de 2 a 6 pb do DNA invasor chamada PAM (*proto-spacer adjacent motif*) (figura 9A,9B), esta sequência, que só existe no DNA alvo e não no DNA do hospedeiro, é reconhecida e auxilia na discriminação entre o proto-espaçador (DNA exógeno) e o espaçador (DNA endógeno) (figura 17), evitando assim a autoimunidade (van der Oost *et al.*, 2014, Marraffini and Sontheimer, 2010a, Shah *et al.*, 2013). O sistema CRISPR tipo III carece da participação da sequência PAM. A discriminação entre o DNA da própria bactéria e o DNA exógeno ocorre pela hibridização da sequência de 8nt das repetições do crRNA com o DNA bacteriano, o que leva a não degradação. Isto não

ocorre entre o crRNA e o DNA exógeno (Marraffini and Sontheimer, 2010b, Tamulaitis *et al.*, 2017) (figura 9C), o que provoca a degradação do alvo.

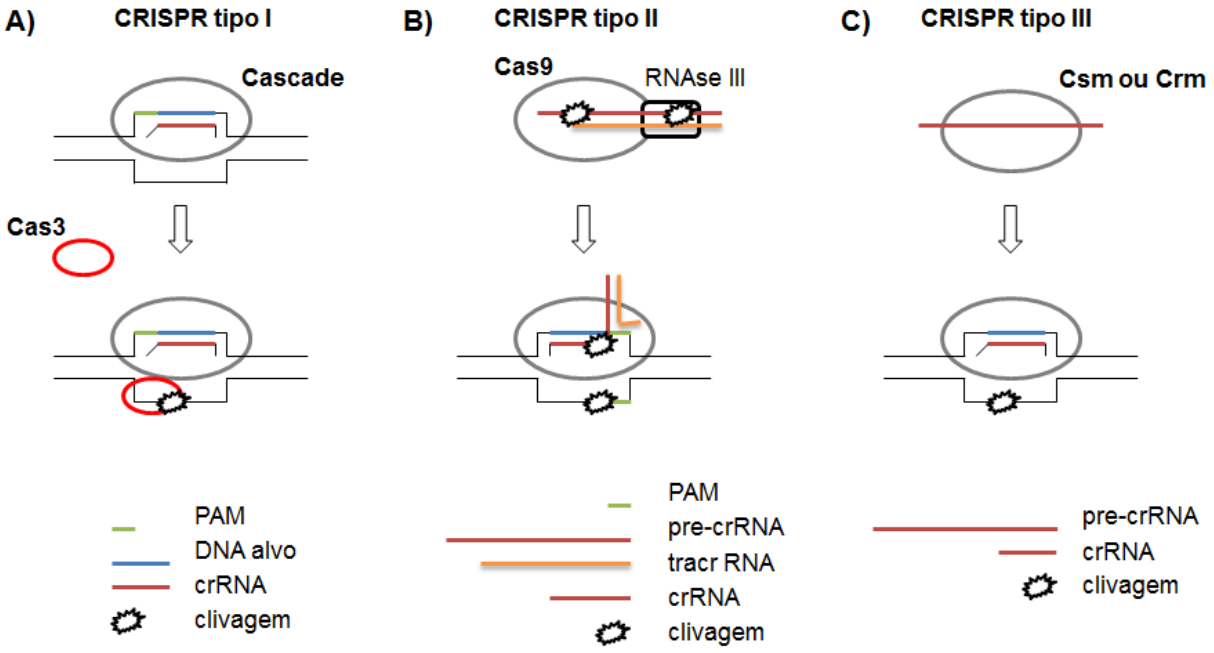


Figura 9. Comparação de mecanismos de ação nos diferentes sistemas CRISPR na presença ou ausência de PAM.

A) O sistema CRISPR tipo I com sequência PAM. B) O sistema CRISPR tipo II (com PAM). C) O sistema CRISPR tipo III (sem PAM).

1.2 Estruturas relevantes do Sistema CRISPR-Cas

Diversos estudos estruturais foram feitos com proteínas isoladas e complexos de RNPs. Publicações de cristalografia de proteínas revelaram em detalhe a composição e função dos complexos crRNP de Cascade (Jore *et al.*, 2011, Wiedenheft *et al.*, 2011a, Jackson *et al.*, 2014, Zhao *et al.*, 2014, Mulepati *et al.*, 2014, Hayes *et al.*, 2016) (figura 10A) e Cmr (Staals *et al.*, 2013, Osawa *et al.*, 2015) (figura 10B). Entretanto, apenas estudos realizados por microscopia eletrônica (de baixa resolução) e espectrometria de massa definiram o complexo crRNP Csm de *Sulfolobus solfataricus* e seu complexo homólogo de *Thermus thermophilus*. Estes estudos indicam que apesar das diferenças de detalhes estruturais, as formas gerais e arquiteturas dos complexos Cascade, Csm e Cmr são muito semelhantes (figura 11), apresentando arquiteturas comparáveis a um “cavalo marinho” (figura 10C).

Este complexo aparenta ser formado por dois filamentos proteicos entrelaçados em forma de hélice. Um destes filamentos foi denominando “filamento dorsal” (em inglês “*backbone*”) e o outro foi denominado “filamento ventral” (“*belly*”), conforme mostrado nas figuras 10A, 10B, 12 e 18. Isto sugere que o complexo efetor de múltiplas subunidades ancestral evoluiu antes da divergência dos sistemas CRISPR-Cas tipo I e III (Makarova *et al.*, 2011b).

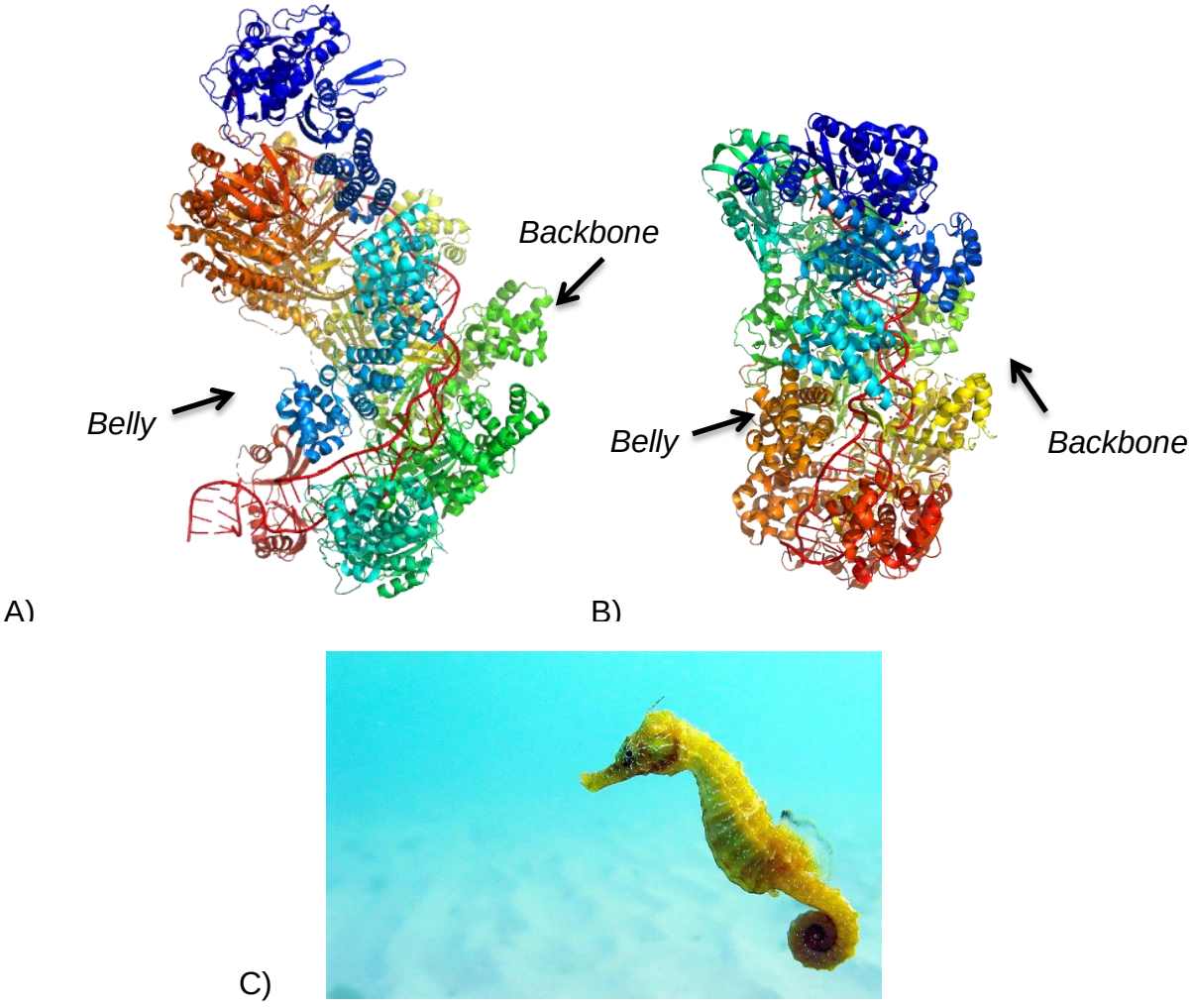


Figura 10. Estruturas relevantes determinadas por cristalografia de proteínas do sistema CRISPR-Cas Classe I que assemelham um cavalo marinho.

A) Complexo Cascade crRNA ligado a DNA alvo fita simples. (organismo: *Escherichia coli*), resolução 3,03 Å PDB: 4QYZ (Mulepati *et al.*, 2014) Esta estrutura permite distinguir as semelhanças comparáveis ao cavalo marinho. B) Estrutura cristalina do complexo tipo III-B CRISPR-Cas e RNA guia ligado a um alvo análogo (quimera proteica construída sinteticamente de proteínas Cas dos organismos *Pyrococcus furiosus* e *Archaeoglobus fulgidus*, resolução 2,09 Å, PDB: 3X1L (Osawa *et al.*, 2015). C) *Hyppocampus guttulatus*, cavalo marinho do Mar Negro, autor Dumitrescu, imagem CC. *Backbone*: Filamento dorsal. *Belly*: Filamento ventral.

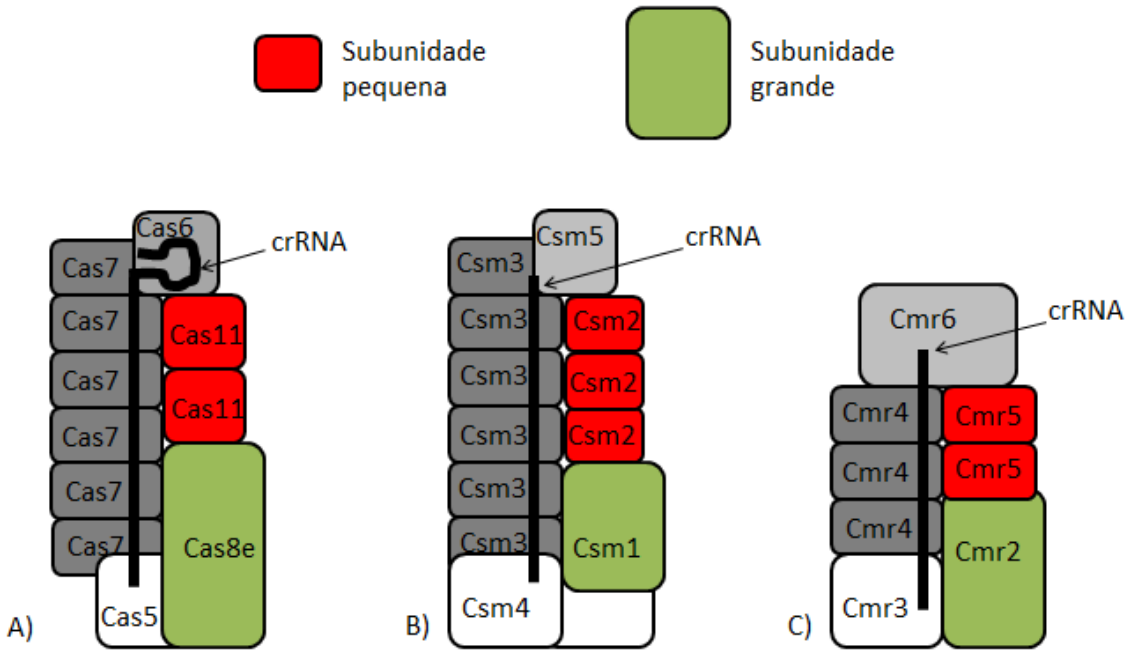


Figura 11. Composição estrutural de complexos crRNP com múltiplas subunidades.

A) Sistema classe I tipo I. B) Sistema classe I tipo III-A (Csm). C) Sistema classe I tipo III-B (Cmr).

1.2.1 Cascade

O sistema tipo I-E de *Escherichia coli* corresponde ao sistema CRISPR mais estudado, ele serve como modelo para o entendimento do funcionamento de outros tipos como é o caso do sistema III-A e III-B devido a suas semelhanças. O complexo proteico característico do sistema CRISPR-Cas subtipo I-E é conhecido como Cascade e está formado por cinco proteínas Cas: Cas8e, Cas11, Cas7, Cas5e e Cas6e (Figura 11A). Estas proteínas possuem diversas nomenclaturas para representar uma mesma proteína. A nomenclatura padronizada e suas terminologias anteriores equivalentes se encontram detalhadas na tabela 1. O Cascade se liga ao crRNA e forma um complexo helicoidal de proteínas RAMP (*repeat-associated mysterious protein*) de 405 kDa. (Mulepati and Bailey, 2013, Mulepati *et al.*, 2014).

Tabela 1. Nomenclatura de proteínas Cas do sistema tipo I-E.

Complexo *Cascade* de *Escherichia coli*		
Nomenclatura atualizada	Equivalente 1	Equivalente 2
Cas8e	CasA	Cse1
Cas11	CasB	Cse2
Cas7	CasC	Cse4
Cas5e	CasD	Cas5
Cas6e	CasE	Cse3

O complexo Cascade é formado por seis unidades formadoras do filamento dorsal do complexo, chamadas Cas7, duas subunidades pequenas Cas11 formadoras do filamento ventral, uma subunidade grande Cas8e, uma proteína Cas5e e uma proteína Cas6e nas extremidades (Mulepati *et al.*, 2014, Jackson *et al.*, 2014, Hayes *et al.*, 2016).

A proteína Cas7 assemelha-se distantemente a estrutura de uma mão, onde a palma da mão liga 5 nucleotídeos do crRNA e o polegar induz uma alça no crRNA (figura 13B). Este padrão é repetitivo levando em consideração as seis proteínas Cas7 (figura 13A e 13B). Já o filamento ventral*,* formado pelas proteínas Cas11 está envolvido na ligação de oligonucleotídeos alvos no complexo Cascade (figura 12C e 12D) (Mulepati *et al.*, 2014, Jackson *et al.*, 2014, Osawa *et al.*, 2015, Hayes *et al.*, 2016).

A proteína Cas8e forma parte da subunidade grande do complexo Cascade (figura 11A). Existem indícios que esta proteína é responsável por identificar a sequência PAM (*protospacer adjacente motif*) e estabilizar o DNA do alvo invasor para desencadear a subsequente degradação do alvo pela endonuclease Cas3 (figura 14) (Sashital *et al.*, 2012, Cass *et al.*, 2015, Hayes *et al.*, 2016). A proteína Cas5e participa como uma proteína de ligação com RNA não catalítica. Esta proteína é responsável pela interação e estabilização estrutural da extremidade 5’ do crRNA (figura 14B e 15). A proteína Cas6e é uma endoribonuclease que processa o crRNA maduro e permanece unida a alça do crRNA na extremidade 3´ (figura 16) (Hochstrasser and Doudna, 2015, Semenova *et al.*, 2015).

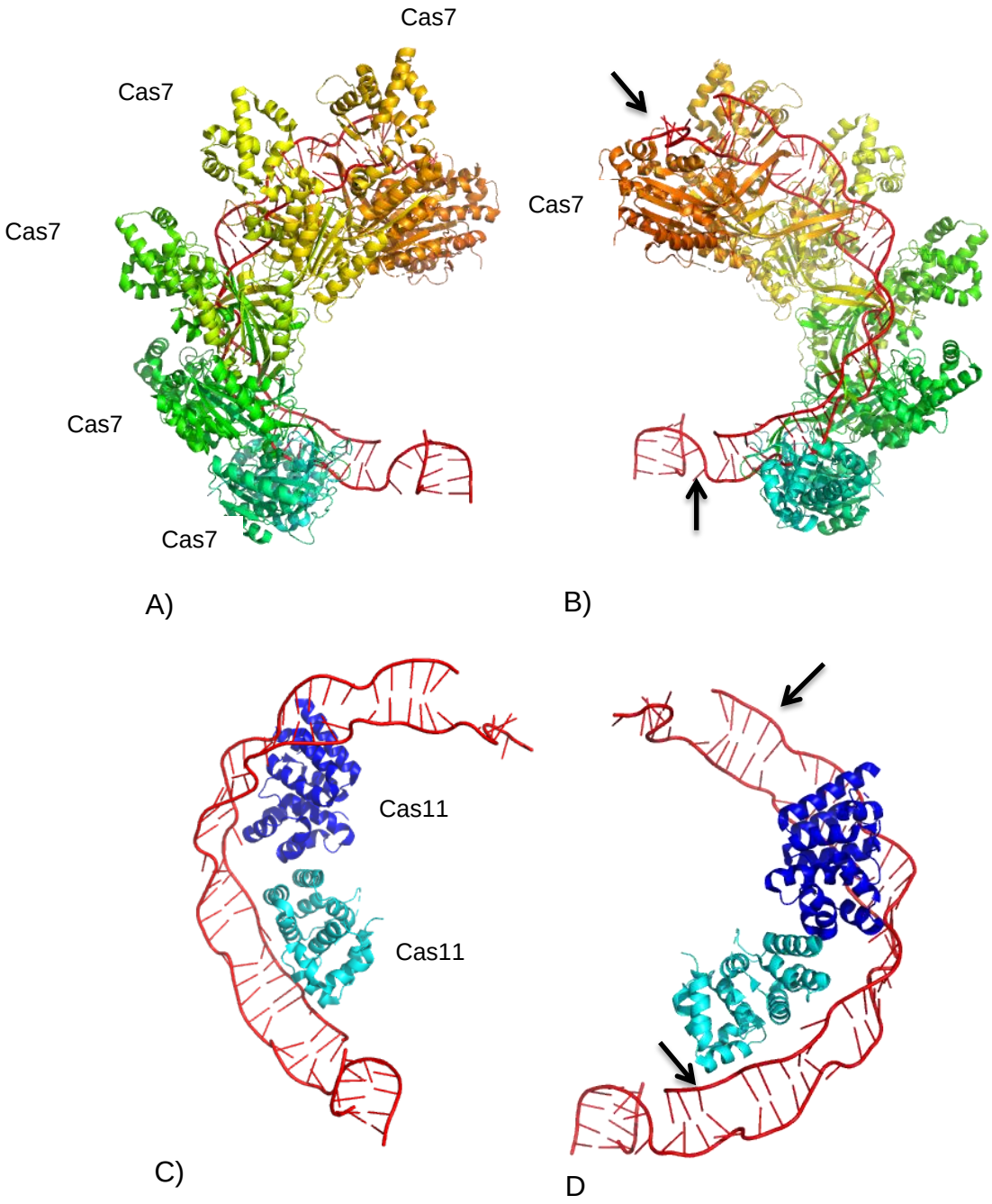


Figura 12. Estrutura parcial do complexo Cascade, junto com um crRNA ligado a seu alvo DNA fita simples.

A) Filamento dorsal do Cascade formado pelas seis proteínas Cas7 que estabilizam o crRNA. B) Imagem de outro ângulo do filamento dorsal, as setas indicam o crRNA. C) O filamento ventral do Cascade formado pelas duas subunidades pequenas Cas11 que estabiliza o DNA fita simples do alvo. D) Imagem de outro ângulo do filamento ventral, mostrando o alvo. Organismo: *Escherichia coli*, Resolução 2,45 Å, PDB: 4QYZ (Mulepati *et al.*, 2014)**.**

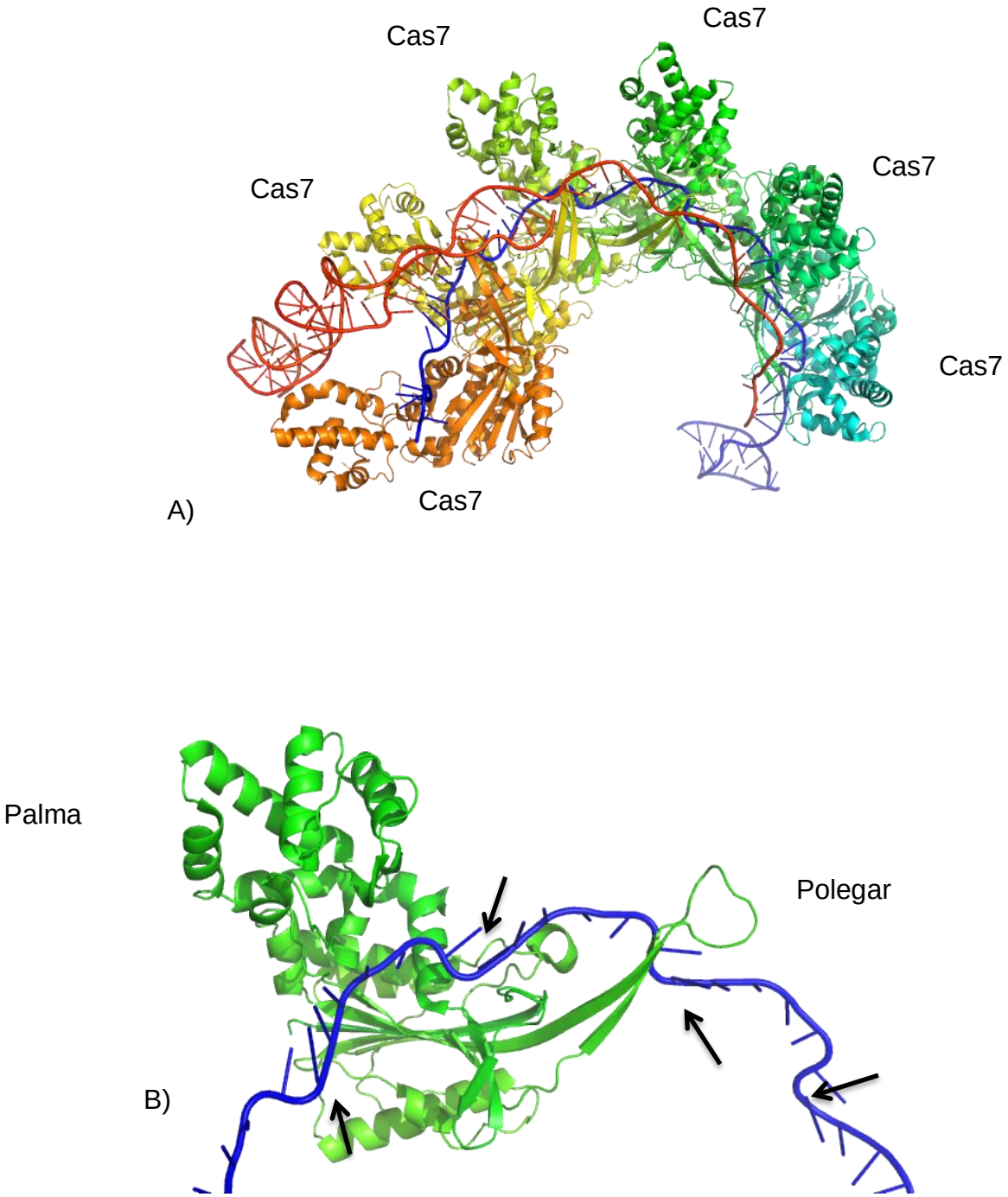
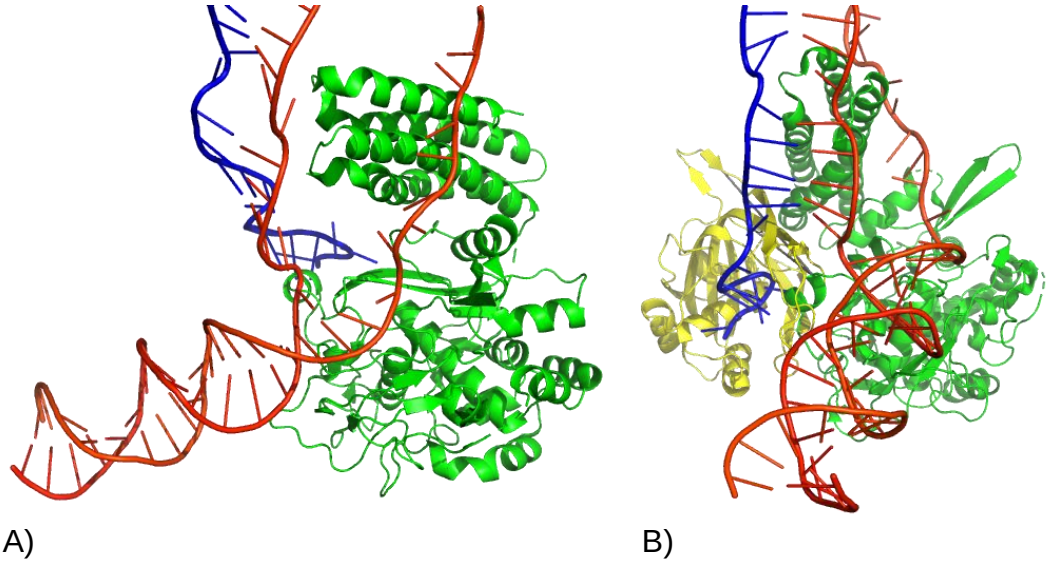


Figura 13. Proteína formadora do filamento dorsal no Cascade.

A) Distribuição das seis proteínas Cas7 formando o corpo vertebral do complexo crRNP junto com seu DNA alvo (vermelho). A proteína Cas7 assemelha uma mão, o conjunto de palmas e polegares da estrutura confere uma forma helicoidal ao complexo. Organismo: *Escherichia coli*. Resolução 2,45 Å, PDB: 5H9F (Hayes *et al.*, 2016). B) O crRNA (azul) mostra a formação de alças repetidas (setas) após um número determinado de nucleotídeos.



A) B)

Figura 14. Proteínas estabilizadoras do DNA alvo e do extremo 5´ do crRNA.

A) Detalhe da proteína Cas8e (subunidade grande em verde) estabilizando o DNA alvo (vermelho). B) Proteína Cas5e (amarelo) e proteína Cas8e (verde) estabilizando crRNA (azul) e DNA alvo (vermelho) respectivamente. Organismo: *Escherichia coli*. Resolução 2,45 Å, PDB: 5H9E. (Hayes *et al.*, 2016).

Figura 15. Proteína Cas5e.

Proteína Cas5e (amarelo) isolada do complexo Cascade ligada ao extremo 5´ do crRNA (azul) e DNA alvo (vermelho). Organismo: *Escherichia coli.* Resolução 2,45 Å, PDB: 5H9F (Hayes *et al.*, 2016).

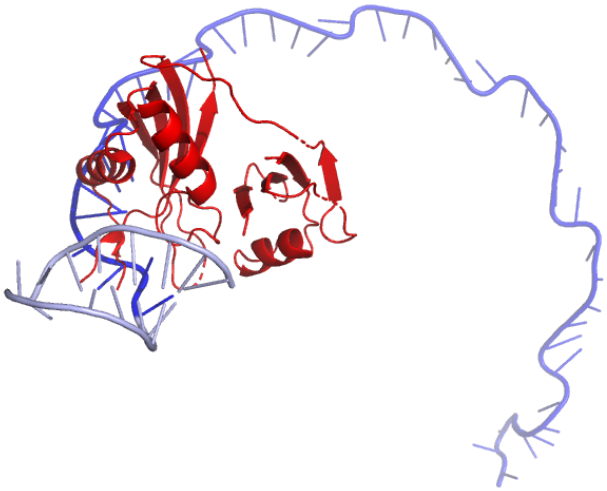


Figura 16. Proteína Cas6e

Proteína Cas6e isolada do complexo Cascade, estabilizando a alça do extremo 3´ do crRNA. (Hayes *et al.*, 2016). Organismo: *Escherichia coli.* Resolução 2,45 Å, PDB: 5H9F.

As proteínas Cas7, Cas5e e Cas6 pertencem a família de proteínas RAMP com motivo RRM (RNA *Recognition Motif*). Estas proteínas possuem alças ricas em glicinas, que ajudam na formação de uma superfície para reconhecimento específico de RNA. (Wang and Li, 2012, Makarova et al., 2011a).

A)

B)

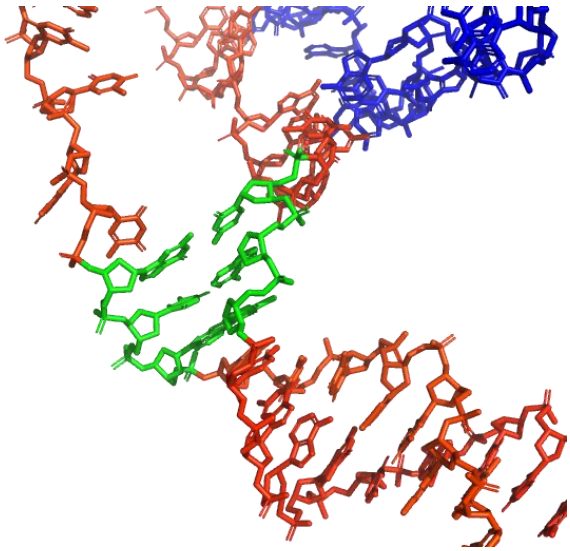


Figura 17. crRNA e DNA alvo.

A) Representação em cartoon do crRNA maduro (laranja), DNA alvo (azul) e sequência PAM (verde). Uma fita simples de DNA se emparelha ao crRNA. A sequência PAM de 3 pb localizado de forma adjacente à sequência alvo B) Representação linear mostrando o detalhe da sequência PAM de 3 pb localizado de forma adjacente à sequência alvo. Resolução 2,45 Å, PDB: 5H9F.

1.2.2 Cmr

A comparação das estruturas cristalográficas em alta resolução dos complexos crRNP do tipo I-E (Cascade) e III-B (Cmr) demonstrou que ambos possuem semelhança estrutural (Mulepati *et al.*, 2014, Jackson *et al.*, 2014, Hayes *et al.*, 2016, Osawa *et al.*, 2015) (figuras 10A e 10B). Como o complexo Cascade, também o complexo crRNP Cmr é formado por um filamento proteico dorsal e um filamento proteico ventral entrelaçados de forma helicoidal (figuras 12 e 18). Como mencionado anteriormente, o filamento dorsal encontra-se envolvido na ligação de crRNA. A função ligante de crRNA de Cas7 no complexo Cascade (figuras 12A e 12B) é realizada pela proteína Cmr4 no complexo Cmr (figuras 18A e 18B). Já o filamento ventral está envolvido na ligação de oligonucleotídeos, sendo que a função da proteína Cas11 no complexo Cascade (figuras 12C e 12D) é realizada pela proteína Cmr5 no complexo Cmr (figuras 18C e 18D) (Mulepati *et al.*, 2014, Jackson *et al.*, 2014, Osawa *et al.*, 2015).

O complexo Cmr apresenta três subunidades Cmr4 e duas subunidades Cmr5 (figura 18). Como a proteína Cas7 de Cascade, a proteína Cmr4 aparenta a forma de uma mão, onde a palma permite ligar 6 nucleotídeos do crRNA e o polegar induz a formação de uma alça no crRNA (figura 19). Uma extremidade dos filamentos helicoidais apresenta a subunidade grande Cmr2 e a proteína Cmr3. A proteína Cmr2 estabiliza o oligonucleotídeo alvo. A proteína Cmr3 reconhece a extremidade 5’ do crRNA e define a posição de partida do duplex guia-alvo. A extremidade 3’ do crRNA interage com a proteína Cmr6 (figura 20).

1.2.3 Cas9

Como discutido anteriormente, no sistema CRISPR-Cas tipo II, o reconhecimento do oligonucleotídeo alvo e a sua clivagem é realizada por uma única proteína, a Cas9. Diversos estudos estruturais descrevem com grande detalhe complexos desta estrutura, inclusive mostrando a conformação do complexo proteína – RNA guia - DNA alvo (Jinek *et al.*, 2014, Nishimasu *et al.*, 2014). Este complexo consiste de dois lóbulos, o lóbulo REC que reconhece o DNA alvo (com domínios REC1 e REC2) e o lóbulo NUC que contém dois domínios com atividade nuclease (RuvC e HNH) para clivagem do DNA alvo e um domínio PI (PAM *interacting domain*). O complexo entre o DNA alvo e um híbrido artificial denomidado sgRNA (*single guide* RNA, uma sequência artificial com sequência crRNA e

tracrRNA), que se encontra na interface entre os dois lóbulos, está indicada na figura 21.

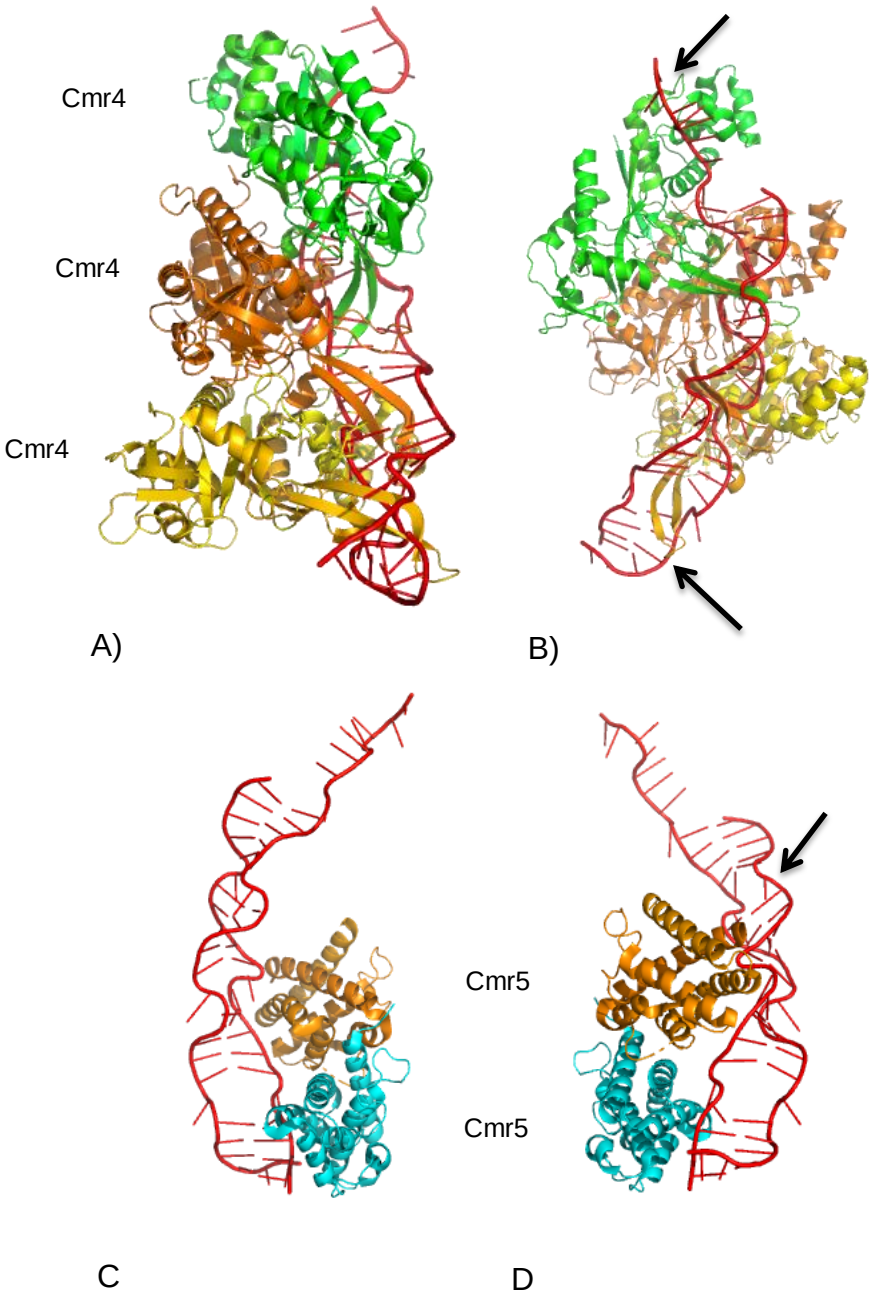


Figura 18. Estrutura parcial do complexo tipo III-B CRISPR Cas junto com um RNA guia ligado a seu alvo análogo.

A) Filamento dorsal do complexo Cmr formado por três proteínas Cmr4 que estabilizam o crRNA. B) Imagem de outro ângulo do filamento dorsal e seu crRNA (indicado por setas). C) Complexo Cmr formado pelas duas subunidades pequenas Cmr5 que estabilizam oligonucleotídeos alvo. D) Imagem de outro ângulo com oligonucleotídeo alvo indicado por seta. Quimera proteica construída sinteticamente por proteínas Cas dos organismos *Pyrococcus furiosus* e *Archaeoglobus fulgidus* (Cmr4/Cmr5), resolução 2,09 Å, PDB: 3X1L (Osawa *et al.*, 2015).

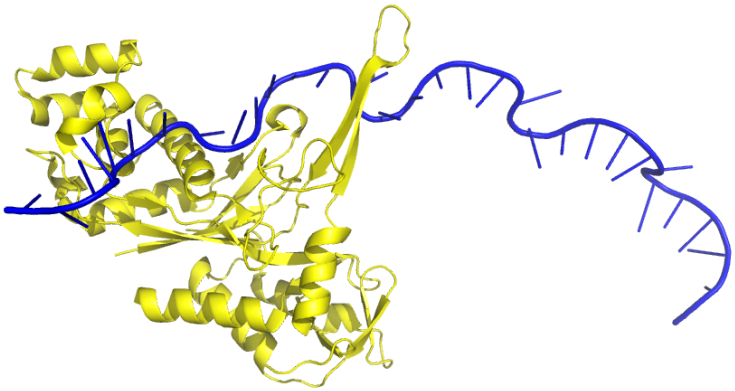


Figura 19. Proteína Cmr4 formando as alças do RNA guia

A proteína Cmr4 assemelha uma mão, o conjunto de palmas e polegares da estrutura confere forma a um filamento helicoidal do complexo Cmr. O crRNA (azul) mostra a formação de alças repetidas a cada 6 nucleotídeos. Quimera proteica construída sinteticamente pelas proteínas Cas dos organismos *Pyrococcus furiosus* e *Archaeoglobus fulgidus* (Cmr4), resolução 2,09 Å, PDB: 3X1L (Osawa *et al.*, 2015).

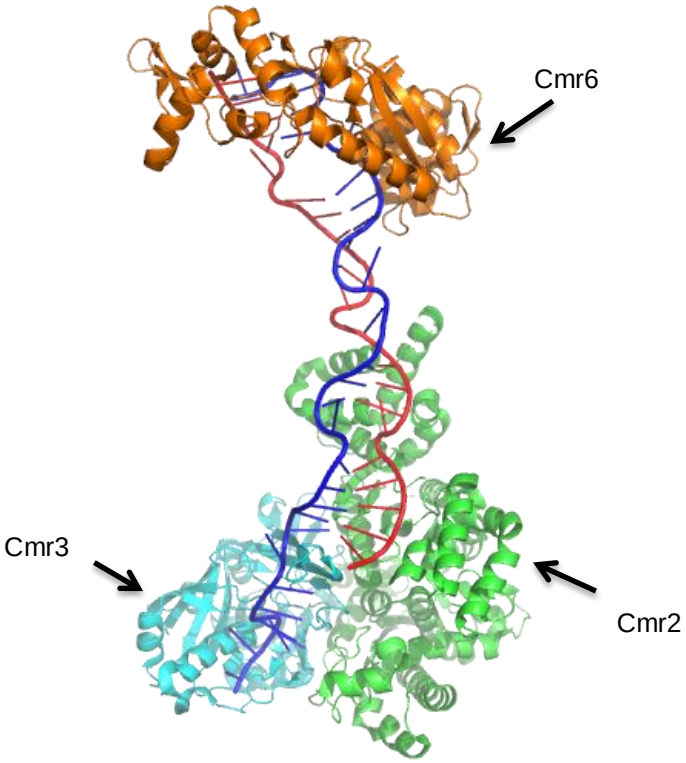


Figura 20. Outras proteínas do complexo Cmr junto com um RNA guia ligado a seu alvo análogo

O heterodímero entre Cmr2 e Cmr3 estabiliza o oligonucleotídeo alvo e o crRNA respectivamente. A proteína Cmr6 (laranja) interage com a extremidade 3´ do crRNA. Quimera proteica construída sinteticamente pelas proteínas Cas dos organismos *Pyrococcus furiosus* (Cmr2/Cmr3) e *Archaeoglobus fulgidus* (Cmr6), resolução 2,09 Å, PDB: 3X1L (Osawa *et al.*, 2015).

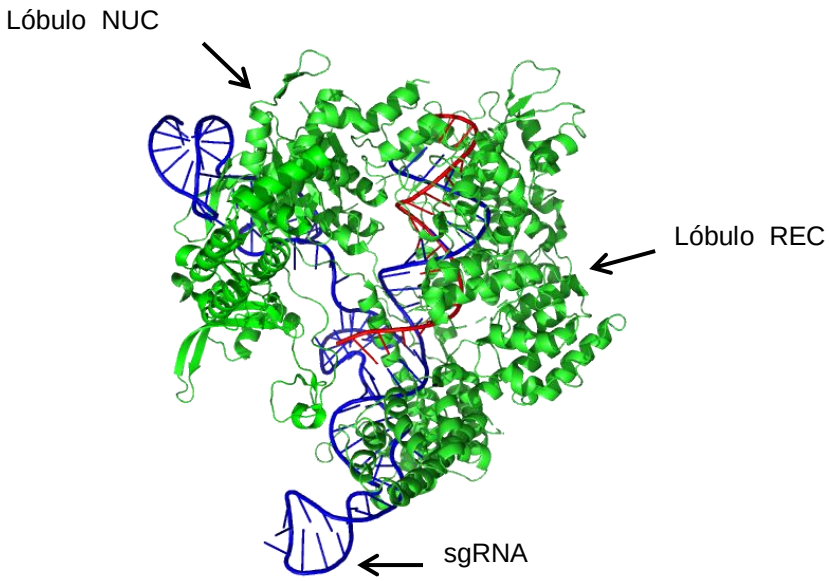


Figura 21. Cas9

Proteína Cas 9 do organismo *Streptococcus pyogenes* em complexo com um RNA guia (azul) e seu alvo DNA (vermelho), resolução de 2,5 Å, PDB: 4OO8 (Nishimasu *et al.*, 2014).

1.2.4 Csm

O complexo Csm (CRISPR-Cas subtipo III-A) tem sido estudado por microscopia eletrônica e espectrometria de massas. O complexo Csm corresponde, portanto, a um sistema menos estudado bioquimicamente e estruturalmente. Originalmente este subsistema ficou conhecido como sistema Mtube por ter sido encontrado na bactéria *Mycobacterium tuberculosis,* que possui somente este tipo de sub-sistema CRISPR-Cas (Groenen *et al.*, 1993, Haft *et al.*, 2005, Jansen *et al.*, 2002).

No caso do complexo Csm de *Sulfolobus solfataricus*, estudos demostraram que o complexo é formado por até sete proteínas (Rouillon *et al.*, 2013). O complexo homólogo de *Thermus thermophilus* consiste aparentemente de cinco proteínas diferentes: a Csm1, Csm2, Csm3, Csm4 e Csm5 (Staals *et al.*, 2014). Apesar da baixa resolução destes estudos com microscopia eletrônica (~30 Å e 17 Å respectivamente), foi possível estabelecer que complexos Csm revelam uma similaridade estrutural global quando comparados aos complexos Cascade e Cmr. Esta semelhança está parcialmente conservada na estrutura e função de suas sub-unidades, como descrito na tabela 2 (Osawa *et al.*, 2015, Rouillon *et al.*, 2013, Koonin *et al.*, 2017).

Tabela 2. Proteínas equivalentes entre três subtipos CRISPR-Cas da classe 1.

Subtipo I-E: Cascade	Subtipo III-A: Csm	Subtipo III-B: Cmr
Cas7	Csm3 – Csm5 (Cas7)	Cmr4 – Cmr6 – Cmr1 (Cas7)
Cas11	Csm2 (Cas11)	Cmr5 (Cas11)
Cas8e	Csm1 (Cas10)	Cmr2 (Cas10)
Cas5e	Csm4	Cmr3
Cas6[1]	Cas6[2]	Cas6[2]

[1] A proteína Cas6 do subtipo I-E e I-F fazem parte do módulo efetor.
[2] A proteína Cas6 tipos III se encontra fora do módulo efetor.

O sistema CRISPR-Cas tipo III-A é formado por seis proteínas Csm3 (figuras 11B e 22C) que formam o filamento dorsal da estrutura (Rouillon *et al.*, 2013, Wiedenheft *et al.*, 2011b, Hatoum-Aslan *et al.*, 2011), da mesma forma como o Cas7 de Cascade forma o filamento dorsal deste complexo (figuras 10A, 12A, 12B e 13A). Tanto a Cas7 como a Csm3 fazem contato com o crRNA (figura 13B) (van der Oost *et al.*, 2014). As proteínas Csm3 são homólogas às proteínas Cas7, por tanto também fazem parte da família RAMP (van der Oost *et al.*, 2014, Hrle *et al.*, 2013). Outras proteínas que também fazem parte da família RAMP são a Csm4 e Csm5. A "cauda" do complexo Csm é formado pela proteína Csm4 que corresponde a proteína Cmr3 no complexo Cmr. A "cabeça" do complexo Csm é formado pela proteína Csm5 que corresponde às proteínas Cmr1 e Cmr6 no complexo Cmr (Spilman *et al.*, 2013, Staals *et al.*, 2013, Rouillon *et al.*, 2013). A proteína Cas8e do complexo Cascade equivale no complexo Csm à proteína Csm1 (conhecida também como Cas10) que forma parte da subunidade grande do complexo (figura 11B). A Csm1 (Cas10) possui um domínio Palm, característico de DNA polimerases e nucleotídeo ciclases que podem ter função estrutural ou função catalítica (Hatoum-Aslan *et al.*, 2014). A Csm1, representada na figura 22A, interage com a proteína Csm4 (figura 22B) no extremo 5´ do crRNA dando estabilidade ao complexo (Hatoum-Aslan *et al.*, 2014) da mesma forma como a Cas8e interage com a proteína Cas5e no complexo Cascade (figura 14B).

Durante o processamento do crRNA no sistema III-A, a proteína Cas6 cliva o pré-crRNA na repetição do extremo 5´, aproximadamente 8 nt antes do começo do espaçador. O tamanho do espaçador varia de 34 a 44 nt e contém 15 a 16nt da seguinte repetição no extremo 3´. A segunda clivagem do crRNA acontece no extremo 3´. A nucleasse responsável por esta clivagem ainda não foi identificada.

Entretanto acredita-se que as proteínas Csm2, Csm3 e Csm5 (todas partes do complexo Csm) são necessárias para a maturação do crRNA. A proteína Csm3 é responsável pelo comprimento final do crRNA para gerar crRNA maduro de 37-45nt (Hatoum-Aslan *et al.*, 2011, Hatoum-Aslan *et al.*, 2013, Hatoum-Aslan *et al.*, 2014).

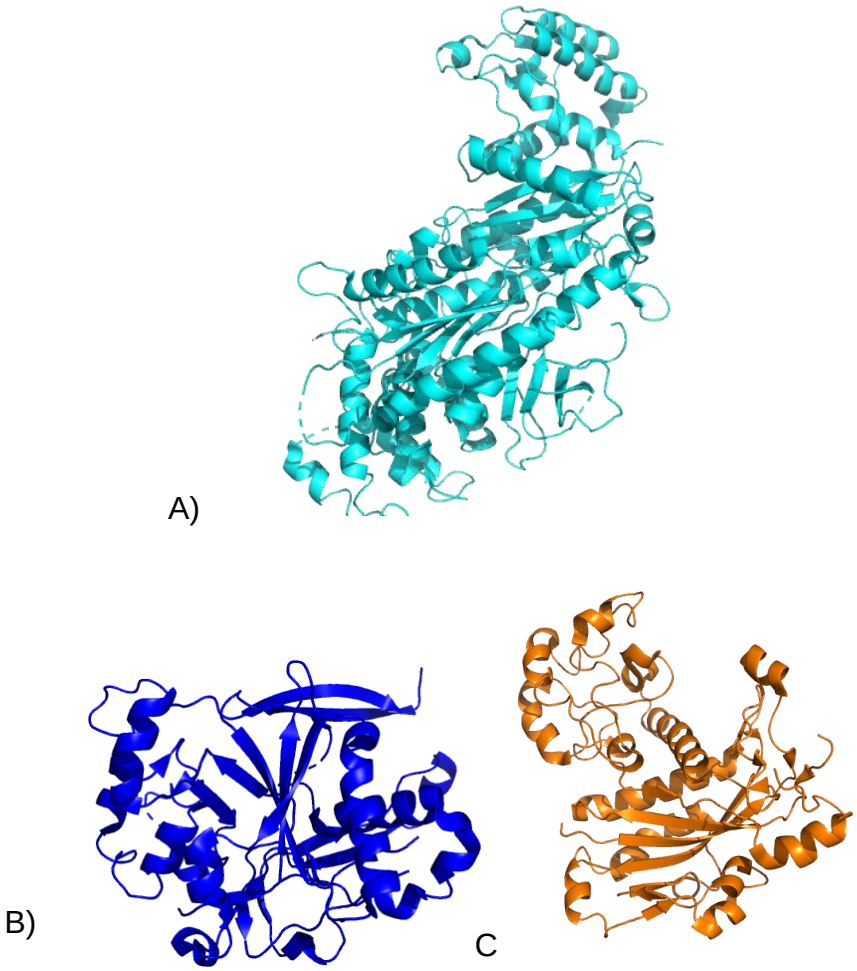


Figura 22. Estruturas de subunidades do complexo crRNP do sistema tipo III-A.

A) Proteína Csm1 do organismo *Thermococcus onnurineus*. PDB: 4UW2.(Jung *et al.*, 2015) B) Proteína Csm4, uma proteína análoga de Cas5e. Organismo *Methanocaldococcus jannaschii.* PDB: 4QTS (Numata *et al.*, 2015). C) Proteína Csm3. Organismo *Methanopyrus kandleri* . PDB: 4N0L (Hrle *et al.*, 2013)

O filamento ventral do complexo Csm é formado pela proteína Csm2, que é o objetivo de estudo deste trabalho. Esta proteína, também denominada de "subunidade pequena" dos sistemas CRISPR-Cas tipo I e III, é formada pelas proteínas Cas11 no subtipo I-E, Csa5 no subtipo I-A e Cmr5 nos subtipos III-B e III-C. No subtipo III-A e III-D esta função aparentemente é atribuída à proteína Csm2. Estudos bioinformáticos indicam que se trata de uma proteína α-helicoidal com

analogia funcional com Cas11, Csa5 e Cmr5 (Reeks *et al.*, 2013a), sendo que uma comparação estrutural entre as subunidades pequenas mostra regiões conservadas entre a proteína Cmr5 e o N-terminal de Cas11, e entre a proteína Csa5 e o C-terminal do domínio Cas11 (Osawa *et al.*, 2015, Jackson *et al.*, 2014, Reeks *et al.*, 2013a, Makarova *et al.*, 2013). Foi ainda sugerido, que a ausência de genes que codificam esta subunidade pequena em determinados óperons Cas (como é o caso dos subtipos I-B, I-C, I-D e I-F) é compensada por extensões maiores de suas respectivas subunidades grandes (Makarova *et al.*, 2011a, Makarova *et al.*, 2013). Por cristalografia de proteínas sabemos que a subunidade pequena do subsistema tipo I-E corresponde a um dímero (Jackson *et al.*, 2014, Mulepati *et al.*, 2014, Hayes *et al.*, 2016). Experimentos realizados por microscopia eletrônica indicam que a subunidade pequena no complexo III-A e III-B, Csm2 e Cmr5 respectivamente, apresentam possivelmente três cópias dentro de seus complexos (Staals *et al.*, 2013, Rouillon *et al.*, 2013, Staals *et al.*, 2014). Entretanto, a determinação da estrutura do complexo Cmr por cristalografia de proteínas, demostrou a formação de um dímero de Cmr5, rejeitando a hipótese de três cópias nesse sistema (Osawa *et al.*, 2015). Devido a sua relevância estrutural para a formação dos complexos Csm, é objetivo deste trabalho, determinar a estrutura cristalográfica da proteína Csm2 de *Thermotoga maritima*.

1.3 THERMOTOGA MARÍTIMA

A *Thermotoga maritima* MSB8 é uma bactéria com propriedade morfológica bacilar e Gram negativa, de natureza anaeróbica e termofílica, que pertence à ordem Thermotogales. Esta bactéria foi isolada de sedimentos marinos geotérmicos da Itália e cresce a uma temperatura entre 60°C a 90°C (Huber, 1986). No microscópio, observa-se que ela se encapsula dentro dum invólucro em forma de bainha, o que faz lembrar uma toga, daí o seu nome (figura 23). As proteínas de organismos termofílicos apresentam, devido a sua estabilidade, uma vantagem importante para aplicações biotecnológicas. Comparado com as proteínas convencionais, que podem desnaturar mais facilmente, as proteínas de organismos termofílicos possuem uma solubilidade maior, expressam por vezes em maiores quantidades e possuem características importantes para a biologia estrutural e aplicações industriais. Por estas razões proteínas de organismos termofílicos podem ser usadas como modelos para estudo de proteínas de outros organismos.

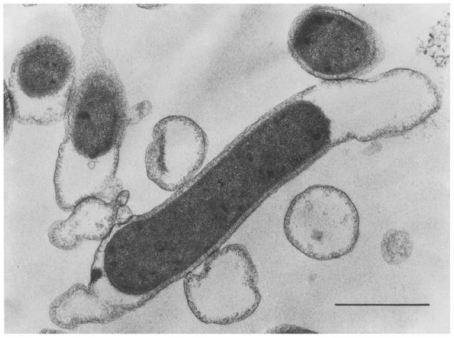


Figura 23. *Thermotoga marítima* MSB8

Bacilo não esporulante Gram-negativo e hipertermófilo. Barra, 1µm (Huber, 1986)

A *Thermotoga maritima* conserva três sistemas CRISPR-Cas que se encontram esquematizados na figura 24: O subtipo III-A ou Csm, o subtipo III-B ou Cmr e outro subtipo menos conhecido chamado Hmari (subtipo I-B) (Haft *et al.*, 2005).

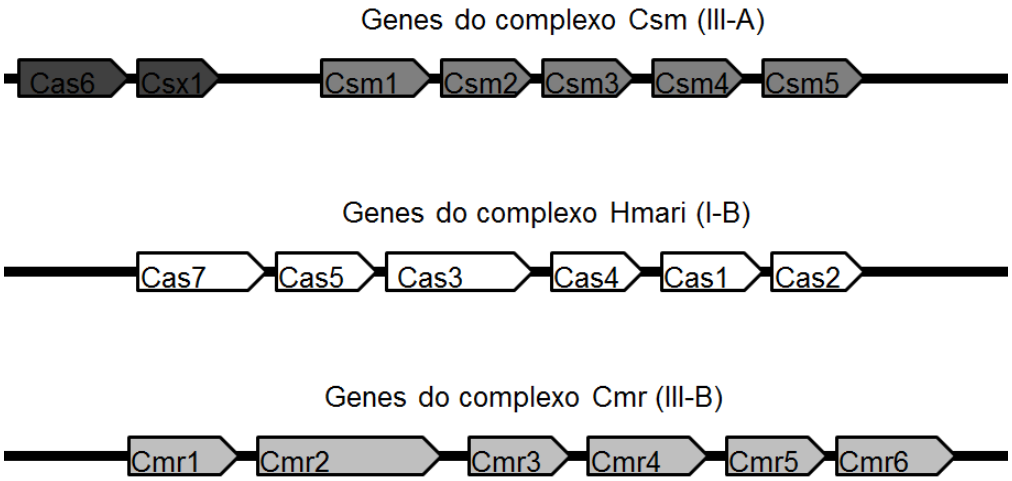


Figura 24. Sistemas CRISPR-Cas encontrados na bactéria *Thermotoga maritima* MSB8.

Subtipo III-A (Csm1, Csm2, Csm3, Csm4 e Csm5), subtipo I-B (Cas7, Cas5, Cas3, Cas4, Cas1 e Cas2) e subtipo III-B (Cmr1, Cmr2, Cmr3, Cmr4, Cmr5 e Cmr6)

1.4 A PRODUÇÃO DE PROTEÍNAS RECOMBINANTES

A tecnologia do DNA recombinante tem sido empregada na produção de diversas proteínas de interesse biológico, por exemplo para a determinação de estruturas macromoleculares, o tratamento de doenças humanas e também na produção de vacinas. As bactérias que carregam um determinado gene clonado em

um plasmídeo são capazes de replicar essa sequência de DNA em múltiplas cópias, bem como de produzir a proteína que esse gene codifica em altas concentrações. Por estas razões as bactérias podem ser utilizadas como verdadeiras indústrias de proteínas. A insulina (Johnson, 1983), hormônio de crescimento (Tritos and Mantzoros, 1998), ou vacina contra a hepatite (Lancaster *et al.*, 1989) são só alguns exemplos para serem citados.

O sistema mais comumente utilizado para expressão de proteínas recombinantes é o que utiliza a bactéria *Escherichia coli* como célula hospedeira. Este sistema é amplamente difundido devido a sua relativa simplicidade, baixo custo, cultivo rápido e pela reprodutibilidade e abundância das proteínas que produz. Sua bem caracterizada genética e o fato de ter à disposição diversas ferramentas para desenvolvimento biotecnológico é um grande atrativo na área da pesquisa básica e aplicada (Sorensen and Mortensen, 2005). Como alternativa podem ser considerados sistemas de expressão em células de eucariotos como a levedura *Pichia pastoris* (Cregg *et al.*, 2000), ou *Saccharomyces cerevisiae* (Evans *et al.*, 2010); ou células de insetos, usando o sistema de baculovírus (Trowitzsch *et al.*, 2010), células de mamíferos (Pollock *et al.*, 1999) ou inclusive plantas e outros organismos complexos.

A expressão de proteínas recombinantes tornou-se uma abordagem de grande relevância que vem revolucionando os estudos de estrutura, função, produção e identificação de novas proteínas.

1.4.1 Vetores para expressão

Um vetor ou plasmídeo para expressão em *Escherichia coli* deve apresentar as seguintes características: Possuir uma origem de replicação, ter um marcador para seleção (por exemplo o gene da β-lactamase que confere resistência a ampicilina) (Goh and Good, 2008), um promotor para transcrição (como o promotor do lac-óperon) e sequências terminadoras de transcrição. Além disso, o vetor deve conter um sistema de repressão para manter muito baixo os níveis basais de expressão do gene até a indução (Williams *et al.*, 1998) o que se faz geralmente pela adição de Isopropil-β-D-1-tiogalactopiranosídeo (IPTG), no caso por exemplo do uso do repressor lac I. Outras características apresentadas num vetor de expressão são: Uma sequência para controle da tradução (por exemplo, um sítio de ligação ao

ribossomo para a iniciação da tradução Shine-Dalgarno e um ATG iniciador), um sinal de terminação da tradução (códon de terminação) também deve estar presente no vetor ou no inserto a ser clonado (Praszkier and Pittard, 2005) e um sítio de múltipla clonagem (MCS) para facilitar a inserção do gene de interesse na orientação correta. Uma vez construído, o vetor de expressão contendo a sequência codificadora da proteína de interesse é introduzido em *Escherichia coli* por transformação.

1.4.2 Produção de proteínas recombinantes

De maneira ideal, quando se pensa em expressão recombinante, espera-se que a proteína de interesse seja estável, não seja tóxica para a bactéria, seja solúvel, seja produzida em grande quantidade e possa ser facilmente purificada.

Um procedimento muito utilizado é o de expressar a proteína de interesse em fusão com uma sequência proteica ou proteína específica que permita a fácil purificação da mesma através de cromatografia por afinidade (Bell *et al.*, 2013). Em geral projeta-se ainda um sítio sensível a uma determinada protease altamente específica como por exemplo a TEV (*Tobacco Etch Virus*), inserido imediatamente após a proteína de fusão, de maneira que a proteína híbrida possa ser clivada liberando a proteína.

É importante estar alerta de que a situação ideal exposta acima nem sempre é atingida. Na realidade, em muitos casos, proteínas produzidas em bactérias de forma recombinante não expressam ou não são solúveis. As vezes as proteínas podem ser tóxicas para a célula (Baneyx, 1999). Algumas proteínas são expressas em baixos níveis. Diversas estratégicas foram desenvolvidas para contornar estes problemas, como trabalhar com diferentes cepas de *E.coli*, modificar condições de expressão, realizar testes em pequena escala, expressar variantes de uma sequência proteica, variar concentração de sal nos tampões de purificação, adicionar glicerol e DTT nestes tampões para evitar respectivamente a precipitação e oxidação da proteína, entre outras estratégias (Graslund *et al.*, 2008).

1.5 CRISTALOGRAFIA DE PROTEÍNAS

A difração de raios-X aplicada à análise de cristais de macromoléculas biológicas é a técnica que mais tem influenciado a bioquímica estrutural, contribuindo para o conhecimento da estrutura tridimensional de proteínas, ácidos nucleicos, vírus e outras macromoléculas. A mioglobina e a hemoglobina foram as primeiras estruturas de proteína determinadas por esta técnica (Kendrew *et al.*, 1958, Perutz *et al.*, 1960). A determinação da estrutura tridimensional de uma proteína por difração de raios-X implica em obter monocristais de alta qualidade da proteína purificada, medir a difração do cristal, utilizar programas computacionais para resolver o problema de fases da cristalografia para calcular e mostrar, por meio de imagens, a densidade eletrônica dentro do cristal e construir um modelo tridimensional da molécula que seja consistente com os dados medidos. Outras técnicas utilizadas para determinar a estrutura de proteínas são a ressonância magnética nuclear (NMR, *Nuclear Magnetic Resonance*), microscopia eletrônica (EM, *Electron Microscopy*), espalhamento de raios X a baixo ângulo (SAXS, Small Angle X-ray Scattering), entre outras.

Em relação à cristalografia de proteínas, está técnica está baseada no fenômeno de difração de raios X por cristais. Em síntese, este fenômeno pode ser descrito através da seguinte transformada de Fourier:

$$\rho(x,y,z) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}|F(hkl)|e^{-2\pi i(hx+ky+lz)+i\alpha(hkl)}$$

Nesta equação $\rho(x,y,z)$ representa a densidade eletrônica do cristal, V o volume da assim denominada célula unitária do cristal (o elemento repetitivo que forma o cristal), $|F(hkl)|$ a amplitude dos assim denominados fatores de estrutura (relacionados com a amplitude das reflexões de raios X do cristal medidos) e α a fase dos fatores de estrutura, sendo que as somatórias são realizadas sobre os índices hkl das reflexões medidos. Uma abordagem mais profunda da técnica de cristalografia é reproduzida em excelentes monografias sobre o assunto (McRee, 1999, Drenth, 2007). Em princípio, os principais gargalos/problemas da cristalografia de proteínas podem ser resumidos em três etapas críticas. 1) Obtenção da proteína pura, estável e em altas concentrações. 2) Obtenção de monocristais com capacidade de difração de pelo menos em torno de 3 Å. 3) Resolução do problema

das fases, ou seja, determinação da fase α de cada fator de estrutura na equação acima.

O terceiro problema se deve às limitações experimentais, que não permitem medir diretamente a fase das reflexões. Este problema denomina se “problema das fases” da cristalografia. Como o conhecimento das fases é primordial para reconstruir a estrutura cristalográfica esta deve ser compensada de alguma maneira. Para resolver este problema, existem várias técnicas que permitem fazer o cálculo das fases para obter a imagem da proteína. Estes métodos são conhecidos como substituição molecular, substituição isomórfica e difração anômala (Taylor, 2010).

O primeiro método corresponde à técnica de substituição molecular (*Molecular Replacement*, MR). Este método utiliza as fases de uma proteína de estrutura definida, servindo de modelo para determinar a estrutura de uma proteína desconhecida porém homóloga ou simplesmente similar estruturalmente(Abergel, 2013).

Se não houver uma estrutura homologa, é preciso utilizar outras técnicas de obtenção das fases experimentais para resolver a estrutura de macromoléculas, como as técnicas de substituição isomórfica (IR, *Isomorphous Replacement*) como substituição isomórfica única (SIR, *Single Isomorphous Replacement*) e/ou substituição isomórfica múltipla (MIR, *Multiple Isomorphous Replacement*). Outras técnicas que podem ser usadas são as técnicas de difração anômala, como a difração anômala múltipla (MAD, Multiwavelength Anomalous Dispersion) e/ou difração anômala simples (SAD, *Single-wavelength Anomalous Dispersion*).

No caso do SIR e do MIR, átomos pesados inseridos no cristal dispersarão os raios-X e produzirão uma diferença na difração com respeito ao cristal nativo (cristal não derivatizado com metal pesado). Esta diferença é utilizada para estimar as fases das reflexões. Para a técnica SIR utiliza-se um tipo de átomo pesado e para o MIR podem ser usados mais de um átomo pesado por cada cristal. Para ambas técnicas é usado um único comprimento de onda e os cristais derivatizados devem ser isomorfos com relação ao cristal nativo (Nagem *et al.*, 2003, Sun *et al.*, 2002).

Já a técnica de difração anômala simples (SAD) permite usar um único conjunto de dados a um único comprimento de onda, diferente da técnica de difração anômala múltiple (MAD), que precisa de vários conjuntos de dados a diferentes comprimentos de onda. Em ambos casos não é necessário dispor de dois cristais ou conjuntos de dados (cristal nativo e cristal derivatizado) como nas

técnicas mencionadas anteriormente (McPherson, 2004, Taylor, 2010), mas sim da incorporação de átomos de elementos com características de difração anômala.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Devido a sua relevância estrutural e funcional para a formação dos complexos RNP como Cascade, Cmr, Csm, é objetivo deste trabalho, caracterizar estruturalmente a proteína Csm2 de *Thermotoga marítima* MSB8.

2.2 OBJETIVOS ESPECÍFICOS

- Clonar o gene correspondente da proteína Csm2 de *T. marítima* MSB8
- Expressar a proteína Csm2 de forma recombinante em *E.coli* BL21(DE3).
- Obter a proteína Csm2 recombinante pura, estável e em quantidades e concentrações suficientes para ensaios de cristalização.
- Determinar a estrutura tridimensional da proteína Csm2 recombinante por difração de raios-X.
- Resolver a estrutura e refinar o modelo na melhor resolução possível.
- Caracterizar bioquimicamente a proteína Csm2.
- Obter da estrutura de Csm2 mais informação sobre a sua função nos complexos Csm.

3 MÉTODOS

3.1 CLONAGEM DO GENE

O gene que codifica a proteína Csm2 a ser clonada encontra-se descrito no banco de dados GenBank do NCBI (*National Center for Biotechnology Information*), este gene identificado pela sigla TM1810, possui o código de acesso AKE29563.1 (GenBank). A sequência de este gene de 422 pb está destacada na figura 25.

```
ATGGCAGTTTCTCAGGGTGTTTCTCTCAAAGAAGATTTGAAGGACCTTGTGAGAAAGGCAGAGGAAATCGGAA
GGGAACTTTCTGGAAAGCTGAAGACAAACCAGCTCAGAAAGTTTCATGGTCACTTAACCAAAATCTGGAGCAA
CTACATCTACAAAAAGAAGGACTACAGGGATAACCCGGAGAAGTTCAACGAAGAGATCCTTAACGAGCTTCAC
TTCATGAAGATATTTCTCGCATATCAGGTTGGAAGGGATATCGAAGGTATCAGTGAATTGAAGGAGATACTTG
AACCTCTCATAGACGAGATAAAGACTCCTGACGAGTTTGAGAAATTCAAAAAGTTCTACGATGCAATCCTTGC
GTATCACAAATTCCATTCTGAATCCGAAAAAAGCAACAGAAGGACAGCCAGAAGATAA
```

Figura 25. Sequência do gene da proteína Csm2 de *Thermotoga maritima* MSB8.

3.1.1 PCR

O gene da proteína Csm2 foi amplificado por PCR usando como molde o DNA da bactéria *Thermotoga maritima* MSB8 (Huber, 1986) obtida via a DSMZ (Coleção alemã de microorganismos e culturas de células, Braunschweig, Alemanha). Os oligonucleotídeos de DNA para a realização da PCR estão descritos na tabela 3. Eles contêm sítios de restrição para as enzimas *Bam*HI e *Hind*III (Exxtend, Paulínia, Brasil).

Na PCR, utilizou-se 50 ng do molde de DNA, 0,2 µM de cada oligonucleotídeo, 0,2 mM dNTP, 1,5 mM de MgCl2 e 2U da enzima Platinum® *Taq* DNA Polimerase (Invitrogen, Carlsbad, EUA) com seu respectivo tampão e água ultrapura q.s.p. 100 µL. O programa de amplificação foi 94°C durante 3 minutos de desnaturação inicial, seguido de 35 ciclos de desnaturação de 94°C por 30 segundos, anelamento de 55°C por 30 segundos e extensão de 72°C por 30 segundos, terminando com uma extensão final de 10 minutos.

Tabela 3. Iniciadores moleculares desenhados na amplificação do gene.

	Sequência	Enzimas
Forward	5´-GGCCGG**GGATCC**GCAGTTTCTCAGGGTGTTTC	BamHI
Reverse	5´-GGCCGG**AAGCTT**ATCTTCTGGCTGTCCTTCTGTTG	HindIII

Em negrito ressaltam as sequências das respectivas enzimas de restrição.

O produto da reação foi verificado em gel de agarose 1% e purificado utilizando o kit QIAquick PCR Purification (Qiagen, Hilden, Alemanha). Após a purificação, realizou-se a dupla restrição dos fragmentos usando as enzimas *Bam*HI e *Hind*III.

3.1.2 Digestão com enzimas de restrição

O fragmento de PCR purificado foi digerido usando as seguintes condições: 50 µL de produto da PCR purificado, 15U da enzima *Bam*HI (Invitrogen, Carlsbad, EUA), 15U da enzima *Hind*III (Invitrogen, Carlsbad, EUA), tampão K (Invitrogen, Carlsbad, EUA) e água ultrapura q.s.p. 60 µL. A reação de restrição foi incubado a 37ºC durante 2 horas.

Ao mesmo tempo, o vetor de expressão pQtev (Protein Struktur Fabrik, Berlin, Alemanha), com código de acesso de GenBank AY243506, foi digerido usando 500 ng de plasmídeo, 30U da enzima *Bam*HI (Invitrogen, Carlsbad, EUA) e 30U da enzima *Hind*III (Invitrogen, Carlsbad, EUA) em tampão K (Invitrogen, Carlsbad, EUA) e água ultrapura q.s.p. 50 µL durante 2 horas a 37°C.

Tanto o produto de PCR como o plasmídeo restringido foram purificados utilizando o kit QIAquick PCR Purification (Qiagen, Hilden, Alemanha). O plasmídeo posteriormente foi desfosforilado usando a enzima SAP (*Shrimp Alkaline Phosphatase*) da seguinte forma: 50 µL do vetor digerido purificado, 2,5U de SAP (USB Corporation, Ohio, EUA), tampão SAP e água ultrapura q.s.p. 60 µL. Esta reação foi incubada durante 2 horas a 37ºC. Posteriormente foi realizada a inativação da enzima mantendo uma temperatura de 65ºC durante 15 minutos.

3.1.3 Ligação

Para a clonagem do inserto, foram utilizados 6,5 µL de produto da PCR clivada e purificada, 2 µL de vetor pQtev linearizado desfosforilado, 1U de T4 DNA ligase (Fermentas, Waltham, EUA) e seu respectivo tampão. A reação de ligação foi incubada a temperatura ambiente durante 2 horas. A figura 24 exibe o mapa do plasmídeo pQtev, permitindo visualizar o sítio de clonagem múltipla (MCS) ou região onde o produto da ligação fica incorporado no vetor. Este sítio é caracterizado por várias sequências de enzimas de restrição, entre as quais se encontram as enzimas *Bam*HI e *Hind*III utilizadas neste estudo.

3.1.4 Transformação

O produto da ligação foi transformado em bactérias *Escherichia coli* TOP10, competentes quimicamente com cloreto de cálcio, através da alternância de temperaturas (4˚C durante 30 minutos, 42˚C durante 2 minutos, 4˚C durante 5 minutos e 37˚C durante 1 hora). Estas bactérias foram plaqueadas em meio LB-ágar com 50 µg/mL de ampicilina devido que o plasmídeo pQtev possui um gene (ß-lactamase) que confere resistência a antibióticos ß -lactâmicos como é o caso da ampicilina (figura 26). As placas foram incubadas dentro de uma estufa bacteriológica a 37°C durante 16 horas. Posteriormente foram selecionadas duas colônias separadamente e inoculadas em meio LB (Luria Bertani: 1% triptona, 1% NaCl e 0,5% extrato de levedura) a 37˚C sob agitação durante um período aproximado de 16 horas.

3.1.5 Extração plasmidial

A extração dos plasmídeos foi realizada utilizando o Qiagen Miniprep Kit® (Qiagen, Hilden, Alemanha) seguindo o protocolo do fabricante. As preparações plasmideais foram submetidas a restrição usando as seguintes condições: 250 ng de plasmídeo, 15U da enzima *Bam*HI e 15U da enzima *Hind*III em tampão K (Invitrogen, Carlsbad, EUA) e água ultrapura q.s.p. 20 µL durante 1 hora a 37°C. As amostras que demonstraram a inserção do clone por eletroforese em gel de agarose 1% foram enviadas para sequenciamento no Laboratório de Biologia Molecular e Diagnóstico Molecular de Doenças Lisossomais (Departamento de Biofísica, UNIFESP).

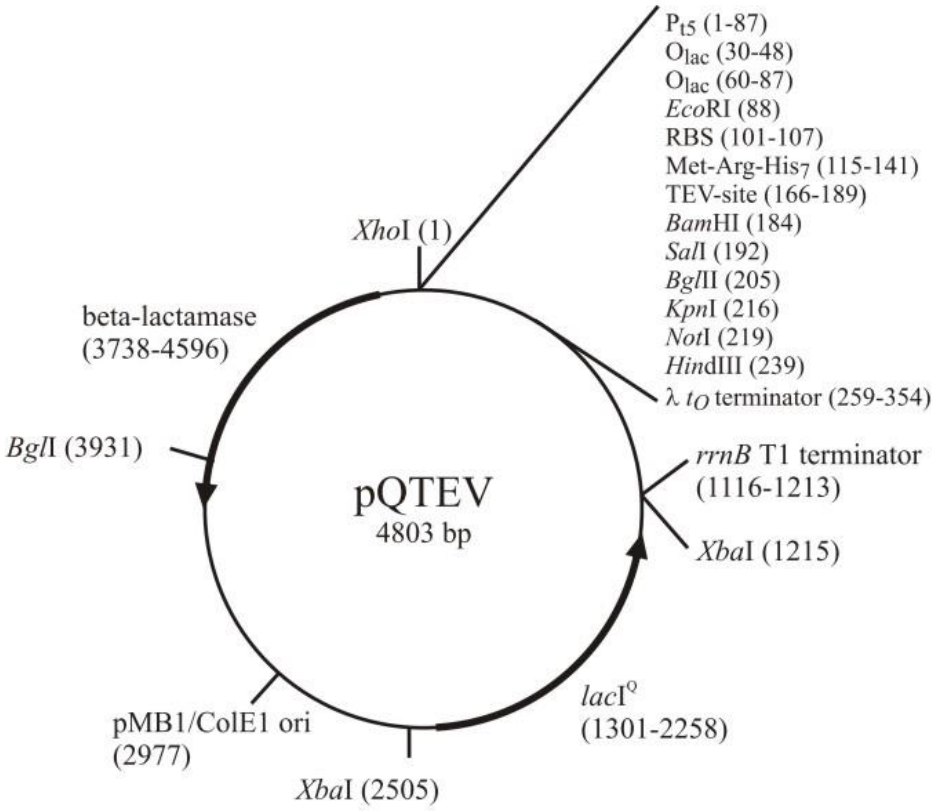


Figura 26. Mapa do vetor pQtev.

O plasmídeo pQtev tem um marcador para seleção (o gene da ß-lactamase que confere resistência a ampicilina), um gene do repressor lacI , um sítio de clonagem múltipla com o promotor da transcrição lac-operon, uma cauda de histidinas no N-terminal, um sítio de clivagem da protease TEV, seis sequências de enzimas de restrição e um códon de terminação entre seus elementos mais relevantes. Imagem disponível no site www.addgene.org/31291/

3.1.6 Eletroforese em gel de agarose

Os géis de agarose 1% foram preparados com tampão TAE (40 mM Tris HCl pH 7,5, 20mM ácido acético e 1mM EDTA). Para visualização das bandas de DNA, foi adicionado prévia solidificação dos géis, e o reagente SYBR® Safe (Invitrogen, Carlsbad, EUA). O tampão utilizado na corrida eletroforética foi o mesmo utilizado na preparação do gel. As eletroforeses foram executadas a 90 V utilizando a fonte Power Pac Basic (Biorad, Hercules, EUA) por 50 minutos aproximadamente. Para visualização de bandas de DNA, os géis foram expostos à luz ultravioleta (comprimento de onda de 302 nm) num transiluminador Benchtop UV Transilluminator (UVP, Analytik Jena, Jena, Alemanha). O SYBR® Safe liga-se ao DNA e forma um complexo que fluoresce quando excitado por radiação UV. Os géis foram fotodocumentados utilizando o PhotoDoc-It Imaging System (UVP, Analytik

Jena, Jena, Alemanha). Todas as amostras de DNA aplicadas em gel foram preparadas adicionando tampão de amostra para DNA (Loading buffer, Invitrogen, Carlsbad, EUA). Nas eletroforeses foram incluídos os marcadores DNA Marker Low Range (Melford, Ipswich, Reino Unido) que possui fragmentos de 250 a 3000pb, 100 bp plus DNA ladder (Bioron, Smart Molecular Solutions, Ludwigshafen, Alemanha), que possui fragmentos de DNA a cada 100 pb e λ DNA/*Hind*III marker (Thermo Fisher Scientific, Waltham, EUA) que possui fragmentos entre 2027 e 23130 pb.

3.1.7 Sequenciamento

O sequenciamento foi realizado utilizando tanto o iniciador molecular pQtev_for como o pQtev_rev, 5′-GAAATTAACTATGAAACATCACCATCAC e **5´-** GGATCTATCAACAGGAGTCC respectivamente, em um aparelho ABI 3130xl Genetic Analyzer (Applied Biosystems, Foster City, EUA) e os resultados foram analisados com o programa Chromas Lite v.2.01 (Technelysium, South Brisbane, Australia).

3.2 EXPRESSÃO DA PROTEÍNA

O gene recém clonado é traduzido numa proteína de 140 aminoácidos e um peso molecular de 16,7 kDa. A sequência de aminoácidos mostrando a tradução do fragmento de interesse correspondendo a proteína Csm2 de *Thermotoga maritima* MSB8 pode ser visualizada na figura 27.

```
MAVSQGVSLKEDLKDLVRKAEEIGRELSGKLKTNQLRKFHGHLTKIWSNYIYKKKDYRDNP
EKFNEEILNELHFMKIFLAYQVGRDIEGISELKEILEPLIDEIKTPDEFEKFKKFYDAILA
YHKFHSESEKSNRRTARR
```

Figura 27. Sequência de aminoácidos codificando a proteína Csm2.

Um plasmídeo com a sequência correta analisada pela ferramenta BLAST – *Basic Local Aligment Search Tool* (NCBI), foi transformado em bactéria de expressão *Escherichia coli* BL21 (DE3). A bactéria posteriormente foi plaqueada em LB-ágar com 50 µg/mL de ampicilina para realizar a expressão da proteína de interesse.

3.2.1 Teste de expressão

Uma colônia de *Escherichia coli* BL21 (DE3) foi selecionada e colocada em meio LB líquido contendo 50 µg/mL de ampicilina. O pré-inóculo foi inserido dentro de um incubador sob agitação a 37˚C durante 16 horas. No dia seguinte o pré-inóculo foi utilizado para inocular 250 mL de meio LB, este meio foi incubado sob agitação a 37˚C até a densidade ótica de 600nm atingir um valor 0,5.

Imediatamente atingida a absorbância requerida, foi retirado uma amostra de 50 mL correspondendo ao momento pre-indução ou 0 horas. Esta amostra foi centrifugada a 4415 g, 4°C, durante 15 minutos. Após centrifugação, o sobrenadante foi descartado e o sedimento foi armazenado a -20°C.

Em paralelo, os 200 mL restantes foram induzidos com 1 mM IPTG (Isopropil β1-tiogalactopiranosídeo) e colocados de volta na incubadora a 140 r.p.m., e 37°C. Após 2, 4, 6 e 16 horas de indução, foram coletadas amostras de 50 mL. Posteriormente estas amostras foram centrifugadas e armazenadas como descrito anteriormente, para a realização subsequente da extração da proteína.

As amostras foram lisadas utilizando um tampão contendo 50 mM Tris-HCl pH 7,5, 100 mM NaCl, 1 mM PMSF, 1% detergente Brij-98, 10 mM imidazol, 5 mM β-mercaptoetanol, lisozima e DNAse. A lise foi incubada durante 20 minutos a 4°C sob agitação. As frações solúveis e insolúveis foram separadas por centrifugação a 21255 g, 4°C, durante 15 minutos.

Aos respectivos sobrenadantes (fração solúvel) de cada amostra foram adicionados 50 µL de micro-grânulos de níquel (Macherey-Nagel, Düren, Alemanha). Estas amostras foram incubadas no gelo sob agitação durante 30 minutos. Logo, foram centrifugadas a 2000 g por 1 minuto e descartado o sobrenadante. Os micro-grânulos sedimentados foram lavados 3 vezes com tampão A (50 mM Tris-HCl pH 7,5, 100 mM NaCl, 10 mM imidazol e 10% glicerol). Os micro-grânulos contendo as frações solúveis foram preparadas para análise em gel SDS-PAGE, desnaturando

as frações com tampão de amostra de proteína e incubando as amostras a 95ºC durante 5 minutos antes da corrida eletroforética.

3.2.2 Expressão em larga escala

Depois de ser realizado o teste de expressão, verificou-se que a proteína Csm2 de *Thermotoga maritima* MSB8 expressa em diversas condições após realizada a indução. A produção foi escalonada a 4 L de meio LB, mantendo a mesma concentração de ampicilina (50 µg/mL) e induzindo a 1 mM de IPTG após a densidade ótica de 600 nm atingir 0,5 como no teste de expressão. Da mesma forma foi mantida a temperatura e agitação de incubação, 37°C e 140 r.p.m. respectivamente. O meio contendo as bactérias foi centrifugado após 6 horas de indução nas seguintes condições: 4415 g, 4°C, durante 30 minutos. O sobrenadante foi descartado e o sedimento bacteriano foi armazenado a -20°C.

3.2.3 Lise bacteriana

O sedimento bacteriano produto da expressão de 4 L em meio LB foi resuspendido com 30 mL de tampão de lise contendo 50 mM Tris–HCl pH 7,5, 400 mM NaCl, 10 mM imidazol, 1 mM PMSF, 5 mM β-mercaptoetanol, 1% Brij 98, DNAse e lisozima. A solução foi incubada sob agitação, a 4°C durante 30 minutos. Posteriormente a amostra foi lisada por pressão mecânica no homogenizador M-110L Microfluidizer (Microfluidics, Newton, EUA) e centrifugada a 21255 g sob refrigeração durante 15 minutos. O sobrenadante foi coletado para posterior purificação.

3.2.4 Purificação

O vetor pQtev agrega uma sequência de 7 histidinas na porção N-terminal da proteína a ser expressa. No total são 24 aminoácidos N terminais agregados, adicionando aproximadamente 3 kDa ao peso molecular da proteína de interesse. Este plasmídeo possui também um sítio de clivagem utilizado pela protease TEV (*tobacco etch virus*). A sequência mostrando as 7 histidinas e o sítio onde é realizada a clivagem pela protease TEV é mostrada na figura 28. A protease

reconhece a sequência de resíduos de aminoácidos ENLYFQG. A clivagem ocorre entre a Glutamina e a Glicina, resultando em uma Glicina na porção amino terminal da proteína de interesse (Fang *et al.*, 2007).

A combinação do plasmídeo pQtev com a bactéria *Escherichia coli* BL21(DE3) facilita a produção de proteínas homólogas com posterior purificação por cromatografia de afinidade em coluna de níquel e seguida de cromatografia de gel-filtração (Sievert *et al.*, 2008).

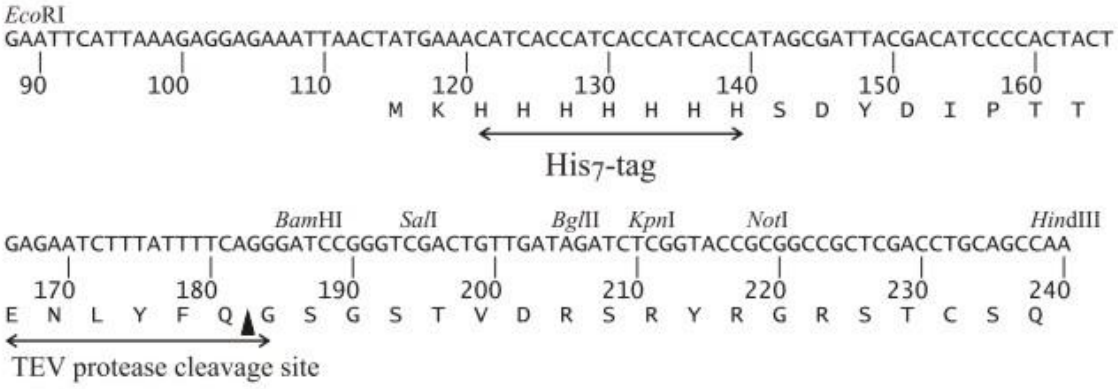


Figura 28. Sequência do plasmídeo pQtev mostrando a cauda de histidinas e o sítio de clivagem da protease TEV.

A sequência de sete histidinas se encontram no N-terminal do sítio de clonagem múltipla do plasmídeo pQtev, seguido do sítio de clivagem da protease TEV reconhecida pela sequência de aminoácidos ENLYFQG e no C-terminal da sequência aparecem as enzimas de restrição *BamHI, SalI, BglII, KpnI, NotI e HindIII* para facilitar a inserção do gene de interesse. Imagem disponível no site www.addgene.org/31291/.

3.2.5 Afinidade a coluna de níquel

Inicialmente foi realizada uma purificação por cromatografia de afinidade usando uma coluna de níquel HisTrap HP de 5 mL (GE Healthcare, Chicago, EUA) através do sistema líquido de cromatografia AKTAprime plus (GE Healthcare, Chicago, EUA). Depois de instalada no aparelho, a coluna foi lavada com água ultrapura e equilibrada com a solução A (50 mM Tris–HCl pH 7,5, 400 mM NaCl, 10 mM imidazol, 1 mM PMSF, 5 mM β-mercaptoetanol e 10% glicerol). O sobrenadante coletado foi injetado no cromatógrafo a um fluxo de 2 mL/min. Antes de eluir a proteína de interesse, a coluna foi lavada novamente com a solução A. Uma solução B contendo 50 mM Tris–HCl pH 7,5, 400 mM NaCl, 500 mM imidazol, 5 mM β-

mercaptoetanol e 10% glicerol, foi utilizada para realizar um gradiente de imidazol com o objetivo de eluir a proteína.

Após este processo, a proteína eluída foi dialisada com tampão C (50 mM Tris-HCl pH 7,5, 400 mM NaCl, 10% glicerol e 3 mM DTT). A diálise aconteceu sob refrigeração durante 4 horas.

3.2.6 Clivagem da cauda de histidinas com a protease TEV

A protease TEV foi produzida de forma recombinante no Laboratório de Bioquímica e Biologia Estrutural do ICT (Instituto de Ciência e Tecnologia) na UNIFESP. A cauda de histidinas foi clivada usando esta protease. A proporção (mg/ml) de TEV:proteína alvo foi de 1:5 utilizando um tampão contendo 1 mM DTT, 0,5 mM EDTA, 50mM Tris HCl pH 8,0. A reação foi incubada sob refrigeração durante 16 horas. Uma subsequente cromatografia de afinidade foi realizada para remover a protease e a cauda de histidinas, utilizando as mesmas soluções A e B descritos na primeira purificação. A proteína clivada foi concentrada até chegar a um volume de 2 mL.

3.2.7 Gel filtração

A proteína concentrada foi injetada novamente no cromatógrafo para separação por gel filtração usando a coluna HiLoad 26/600 Superdex 75 prep grade) GE Healthcare, Chicago, EUA) em tampão D (25 mM Tris-HCl pH 7,5, 100 mM NaCl, 3 mM DTT e 10% glicerol) usando um fluxo de 1 mL/min. A amostra foi logo concentrada dentro de um filtro de centrifugação Amicon Ultra-15 (Millipore, Burlington, EUA) até atingir uma concentração de 15 mg/mL.

3.2.8 Dosagem das proteínas

As frações eluídas mostraram picos de absorbância de 280nm significativos, as quais foram coletadas e dosadas utilizando uma curva padrão de albumina de soro bovino e o reagente Bradford (Serva Electrophoresis, Heidelberg, Alemanha). O comprimento de onda utilizado a medir a absorbância foi de 595 nm.

3.2.9 Eletroforese em gel de poliacrilamida

Em cada passo da produção da proteína, foi confirmada sua presença através de eletroforese em gel de poliacrilamida desnaturante (SDS-PAGE) de 12%. A camada do gel denominado separador consiste em 12% mix de acrilamida/bisacrilamida (29:1), 375 mM Tris HCl pH 8,8, 0,04% TEMED e 0,1% persulfato de amônio. O gel concentrador consiste em 5% mix de acrilamida/bisacrilamida (29:1), 125 mM TRIS-HCl pH 6,8, 10% SDS, 0,1% TEMED e 0,1% persulfato de amônio). Nas corridas eletroforéticas foi utilizado uma solução contendo 25 mM Tris, 192 mM glicina e 0,1% SDS. As amostras de proteína preparadas com o respectivo tampão de amostra para SDS-page e posteriormente aquecidas durante 5 minutos a 95°C. O marcador de peso molecular utilizado foi o BenchMark™ Protein Ladder (Invitrogen, Carlsbad, EUA) mostrando proteínas de 6 a 180 kDa e SeeBluePlus2 (Invitrogen, Carlsbad, EUA) mostrando proteínas de 6 a 195 kDa e LMW Amersham™ (GE Healthcare, Chicago, EUA) mostrando proteínas de 14,4 a 97 kDa. O equipamento de eletroforese utilizado foi Mini PROTEAN Tetra Cell da (Biorad, Hercules, EUA) e as corridas eletroforéticas foram executadas a 50 mA por gel utilizando a fonte Power Pac Basic (Biorad, Hercules, EUA) por 60 minutos aproximadamente. Os géis foram corados com uma solução corante (0,1% Coomasie Blue G-250, 25% etanol e 5% ácido acético) e posteriormente descorados numa solução de 5% ácido acético e 10% etanol para certificar as massas das proteínas purificadas.

3.3 CRISTALIZAÇÃO DA PROTEÍNA

3.3.1 Varredura de cristalização da proteína

A proteína Csm2 foi submetida a ensaios de cristalização, utilizando uma varredura de condições de três kits de cristalização de proteínas (Hampton Research, Aliso Viejo, EUA). Foram usados os kits Crystal Screen, Crystal Screen 2 e o PEG/Ion Screen. A técnica de cristalização utilizada foi a gota suspensa ou *hanging drop* (figura 29) em placas de 24 poços VDX (Hampton Research, Aliso Viejo, EUA). A cada placa foi aplicado silicone (Hampton Research, Aliso Viejo, EUA) nas bordas superiores de cada poço.

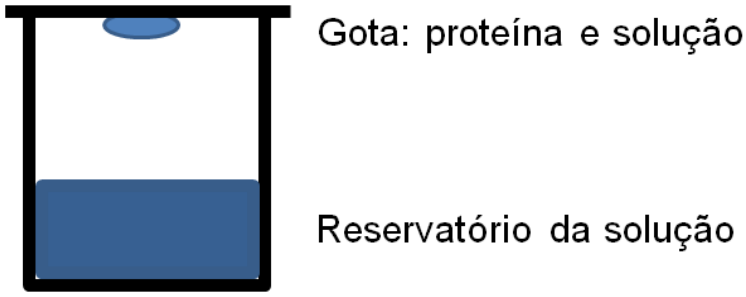


Figura 29. Técnica hanging drop de cristalização.

Esta técnica deixa a gota (proteína e solução) suspensa na lamínula.

Para cada poço da placa foram aplicados 500 µL das soluções dos kits. Em lamínulas de 18 x 18 mm (Knittel Glass, Braunschweig, Alemanha) foram colocadas uma gota de proteína e uma gota de solução do poço reservatório correspondente. As lamínulas foram invertidas sobre cada poço. O sistema é vedado graças ao silicone para poder evitar a evaporação e promover a saturação do sistema. Todas as placas foram armazenadas dentro da uma incubadora a 20°C.

3.3.2 Refinamento da cristalização da proteína

Depois de definir as condições de obtenção de cristais de proteína, é preciso realizar uma otimização das condições de cristalização variando gradativamente a concentração do precipitante, pH e a concentração de sal. Este refinamento ajuda na formação de melhores cristais de proteínas (monocristais em três dimensões de bordas regulares e usualmente cristais maiores). A tabela 4 descreve os dados relevantes sobre a cristalização da proteína Csm2, mostrando a condição 12 do kit Crystal Screen 2 otimizado para 100 mM acetato de sódio pH 4,6, 100 mM CdCl2 e 21% PEG 400.

Tabela 4. Informação sobre a cristalização da proteína Csm2.

Método	***Hanging drop* (gota suspensa)**
Placa	VDX de 24 poços
Temperatura de incubação	20°C
Concentração da proteína	15 mg/mL
Composição do tampão da proteína	25 mM Tris HCl pH 7,5, 100 mM NaCl, 3 mM DTT e 10% glicerol
Condição de cristalização	Crystal Screen 2, condição 12
Composição do reservatório da solução inicial	100 mM acetato de sódio pH 4,6, 100 mM CdCl2 e 30% PEG 400
Composição do reservatório após refinamento	100 mM acetato de sódio pH 4,6, 100 mM CdCl2 e 21% PEG 400
Volume e proporção da gota	4 µL, 1:1 (proteína:solução)
Volume do reservatório durante as otimizações	1 mL

3.3.3 Coleta e processamento de dados cristalográficos

Os cristais de proteína foram medidos no Laboratório Nacional de Luz Síncrotron (LNLS, Campinas, Brasil) tanto na linha de luz MX1 (Polikarpov *et al.*, 1998) como na linha MX2 (Guimaraes *et al.*, 2009). Antes da coleta de dados, os cristais foram transferidos por alguns segundos para uma solução crioprotetora contendo a condição de cristalização suplementada com 12% de glicerol. Os cristais foram coletados da gota utilizando-se uma alça de nylon de 0,2 - 0,4 mm (Hampton Research, Aliso Viejo, EUA). Durante a coleta de dados, cada cristal foi resfriado a -173°C com nitrogênio gasoso para evitar a degradação do cristal de proteína. Os dados cristalográficos foram coletados a um comprimento de onda de 1,458 Å. Estes dados foram analisados usando o software XDS (Kabsch, 2010) e Adxv (Arvai, 2015). O cálculo do coeficiente de Matthews (Vm) foi realizado usando o módulo Xtriage do programa Phenix (Adams *et al.*, 2010). Este coeficiente permite fazer a análise do conteúdo da célula unitária do cristal, determinando o número de subunidades e porcentagem de solvente na unidade assimétrica. O Vm indica a razão entre o volume da célula unitária e o peso molecular das amostras cristalizadas, variando entre 1,62 e 3,53 $Å^3$/Da. Assim, pela comparação de valores obtidos nos experimentos de difração de raios-X com os valores previstos por Matthews, é possível estimar o número de moléculas na unidade assimétrica.

3.3.4 Determinação da estrutura

A estrutura da proteína Csm2 de *Thermotoga maritima* foi solucionada pelo método de difração anômala simples (SAD) usando o software Phaser (McCoy *et al.*, 2007) no módulo Autosol do programa PHENIX (Adams *et al.*, 2010), e aproveitando o fato que as condições de cristalização contém o íon cádmio. O software Coot (Emsley *et al.*, 2010) permitiu inspecionar os mapas de densidade eletrônica. As figuras foram geradas usando o programa Pymol (DeLano, 2015).

3.3.5 Refinamento

O mapa de densidade eletrônica foi inspecionado usando o programa Coot (Emsley *et al.*, 2010). Este mesmo programa foi utilizado para construir o modelo estrutural da proteína Csm2. O refinamento foi realizado com os programas CNS (Brunger *et al.*, 1998, Brunger, 2007) e phenix.refine (Afonine *et al.*, 2012). A validação das estruturas foi realizada utilizando o programa MolProbity implementado no programa PHENIX. As imagens finais foram geradas usando o programa Pymol. Como principal fator da qualidade refinamento do modelo final figuram os fatores cristalográficos R, definidos pela seguinte equação:

$$R = \frac{\sum_{hkl} ||F_O| - k|F_C||}{\sum_{hkl}|F_O|}$$

Nesta equação $\sum hkl$ é a somatória dos índices hkl dos reflexos medidos, k é um fator de escalamento, $|F_O|$ são as amplitudes dos fatores estruturais observados e $|F_C|$ são as amplitudes dos fatores estruturais calculados do modelo. Geralmente estruturas cristalográficas são caracterizadas por dois fatores R, o R*work* e o R*free*. O Rfree é obtido de forma análogo ao Rwork com um subconjunto dos reflexos (geralmente em torno de 10%) que não são usados no refinamento, para assim evitar a introdução de um viés das fases do modelo.

3.3.6 Análise estrutural

Uma vez definida a estrutura, foi calculada o potencial eletroestático da proteína usando o programa APBS (Baker *et al.*, 2001) para visualizar a distribuição

de cargas eletroestáticas em torno dela. Outro modelo calculado baseado com os dados estruturais foi o cálculo de superfície de área, realizado com o programa PDBePISA (Krissinel and Henrick, 2007).

Na busca de homólogos de estruturas para Csm2 foi realizado uma análise comparativa de sequências entre *T. maritima, S. solfataricus* e *T. thermophilus* usando o programa CLUSTALΩ (Sievers *et al.*, 2011).

3.4 CARACTERIZAÇÃO BIOQUÍMICA DE CSM2

3.4.1 Determinação do estado de oligomerização

Os estados de oligomerização da proteína recombinante foram avaliados por cromatografia de exclusão molecular utilizando uma coluna HiLoad 26/600 Superdex 75 prep grade (GE Healthcare, Chicago, EUA), a qual foi equilibrada em tampão D (25 mM Tris-HCl pH 7,5, 100 mM NaCl, 3 mM DTT e 10% glicerol) usando um fluxo de 1 mL/min.

3.4.2 Espectrometria de massas

A massa da proteína nativa Csm2 foi confirmada por espectrometria de massas com ionização por *electrospray* usando um sistema quadrupolo híbrido (Q)-IM-ToF (Synapt G2 HDMS mass spectrometer, Waters). As amostras foram analisadas pelo Prof. Dr. André Zelanis do ICT da UNIFESP no Laboratório de Espectrometria de Massas do LETA-Cetics-Instituto Butantan, São Paulo. A deconvolução do espectro de proteína coletou uma massa molecular média isotópica que assemelha a resultados observados por SDS-PAGE (16732,78 Da). A massa molecular da proteína nativa Csm2 das frações monoméricas e multiméricas derivadas da cromatografia de exclusão molecular foram confirmadas usando o espectrômetro de massa LTQ-Orbitrap Velos (Thermo Fischer Scientific, Waltham, MS, EUA). Os gráficos do espectrómetro de massas foram providenciados usando o programa Orbitrap Analyser. A resolução adquirida nos gráficos foi de 100000 (m/z 400). A realização de uma cromatografia líquida de nanofluxo foi efetuada no equipamento Easy nLC nanoHPLC (Thermo Fischer Scientific) acoplado ao espectrômetro de massa. As frações de proteína foram aplicadas na coluna em

tampão de ácido fórmico a 0,1%, as amostras foram eluídas a um gradiente de 40 minutos aplicando de 0% a 85% de um tampão contendo acetonitrila e 0,1% de ácido fórmico. O espectro de massas de proteínas foi visualizado usando o módulo Qual Browser do programa XCALIBUR (Thermo Fischer Scientific). A deconvolução do estado de cargas foi realizada usando varreduras de espectrometria de massas de média 50-200 para cada corrida LC-MS.

4 RESULTADOS

4.1 CLONAGEM DO GENE CSM2

A amplificação via PCR do gene que codifica a proteína Csm2 foi realizada usando como molde o DNA da bactéria *Thermotoga maritima* MSB8. Após otimização das condições de PCR, uma única banda de aproximadamente 400pb (o gene Csm2 possui 422pb) foi obtida. Esta banda apareceu em quase todas as reações como evidenciado na figura 30.

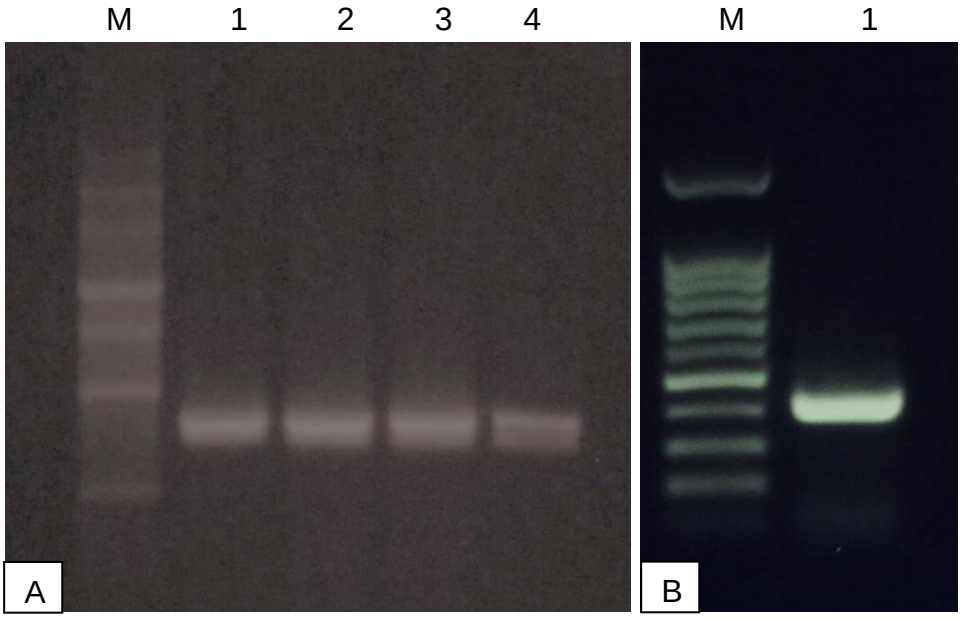


Figura 30. Amplificação do gene Csm2 de 422pb a partir do DNA de *Thermotoga maritima* MSB8.

Géis de agarose de 1%. A) Eletroforese de PCR em gradiente a 60, 55, 50 e 45°C nas canaletas 1, 2, 3 e 4 respectivamente. Marcador de DNA Low Range (Melford, Ipswich, Reino Unido) mostrando bandas de 250, 500, 750, 1000, 1500, 2000 e 3000pb na canaleta M B) PCR realizada a 55°C na canaleta 1. Marcador 100pb plus DNA ladder (Bioron, Ludwigshafen, Alemanha) bandas a cada 100pb na canaleta M.

A primeira PCR realizada foi usando um gradiente decrescente de 60°C a 45°C (figura 30A) para determinar a temperatura ideal de anelamento dos iniciadores (*primers*) moleculares com o molde de DNA. Temperaturas acima de 50°C mostraram uma banda única, indicando que nesta temperatura os iniciadores se anelaram corretamente ao molde evitando ocorrência de hibridizações inespecíficas. A 45°C é possível observar a formação de duas bandas próximas o que pode evidenciar um anelamento inespecífico. Na figura 30B é mostrada outra amplificação do gene de Csm2, de esta vez selecionando exclusivamente a temperatura de anelamento de 55°C.

Para clonar o gene de Csm2 no plasmídeo de expressão bacteriano pQtev, utilizamos os sítios de restrição *Bam*HI e *Hind*III no sítio de clonagem múltipla do vetor. Nas extremidades do produto da PCR foram introduzidos estes mesmos sítios de restrição. Para este fim, os iniciadores moleculares foram extendidos nos seus extremos 5´ com as sequências *Bam*HI para o iniciador *forward* e *Hind*III para o iniciador *reverse.*

Uma forma de prevenir a autoligação do plasmídeo digerido foi adicionar fosfatase alcalina de camarão (SAP) para desfosforilar os extremos do plasmídeo linearizado. A figura 31 mostra o plasmídeo digerido com as enzimas de restrição *Bam*HI e *Hind*III desfosforilado.

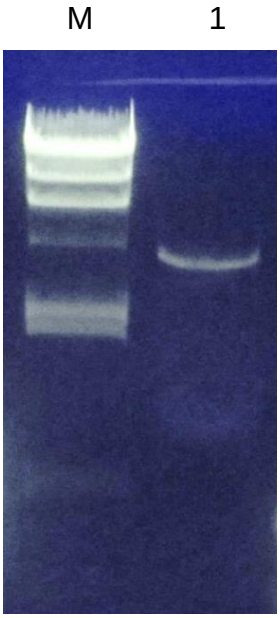


Figura 31. Digestão do plasmídeo pQtev com enzimas BamHI e HindIII. Gel de agarose de 1%.

O plasmídeo pQtev linearizado (canaleta 1) e marcador λ DNA/HindIII marker (Thermo Fisher Scientific, Waltham, EUA) mostrando bandas de 2627, 2322, 4361, 6557, 9416 e 23130pb na canaleta M.

Após realizada a digestão plasmidial e a digestão do produto da PCR, foi executada a ligação do produto PCR ao vetor. O vetor resultante foi em seguida transformado em bactérias competentes *Escherichia coli* TOP10, cepa utilizada em clonagens.

O resultado da clonagem foi evidenciado após extração plasmidial de duas colônias bacterianas e sua respectiva digestão com as enzimas *Bam*HI e *Hind*III. A verificação da clonagem pode ser observada na figura 32. O fato de aparecer, em

duas bandas na última canaleta do gel de agarose nesta imagem indica que a clonagem foi bem-sucedida.

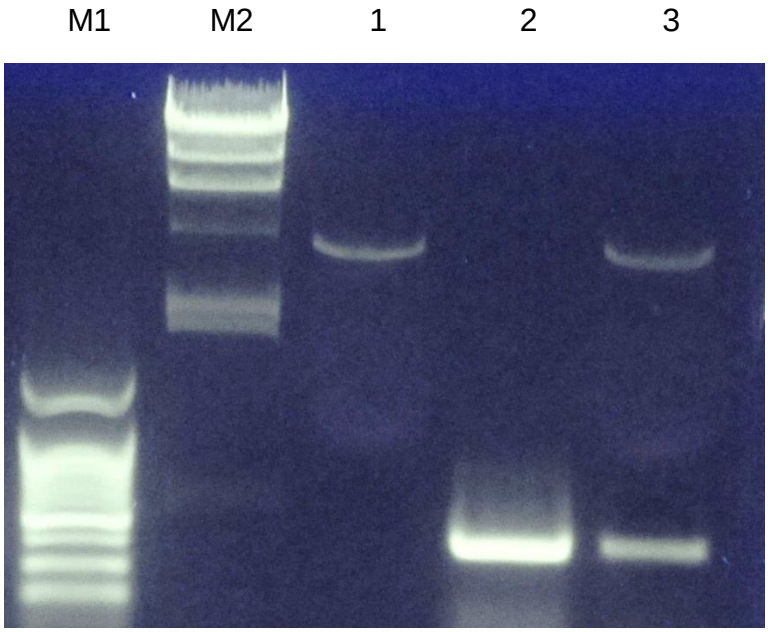


Figura 32. Restrição das extrações plasmideais com as enzimas BamHI e HindIII.

Gel de agarose de 1% mostrando o marcador de 100pb (canaleta M1) e o marcador λ DNA/HindIII (canaleta M2), seguido da preparação plasmidial digerida sem o inserto (canaleta 1), gene Csm2 amplificado por PCR (canaleta 2) e preparação plasmidial com o inserto (banda inferior da canaleta 3).

Para conferir a clonagem, a construção obtida foi adicionalmente sequenciada. A ferramenta computacional BLAST da sequência clonada com o banco de dados do NCBI revelou 100 % de sobreposição com o gene utilizado como referência para clonagem. Este clone após realizado o sequenciamento e verificação da sequência sem ocorrências de mutações, foi transformado na cepa de expressão, *Escherichia coli* BL21(DE3).

4.2 EXPRESSÃO DE CSM2 RECOMBINANTE

Após transformação em *E.coli* BL21(DE3), o clone foi induzido durante a fermentação com 1mM de IPTG e incubado sob agitação por 2, 4, 6 e 16 horas. Foi possível detectar uma banda correspondendo ao peso esperado da proteína Csm2 a partir de 2 horas de indução em todas as frações solúveis, mostrando mais intensidade na fração correspondente a 6 horas de indução. A figura 33 mostra a

expressão de Csm2 recombinante a diversas horas de indução, mantendo uma agitação constante de 140 rpm e incubada a 37°C .

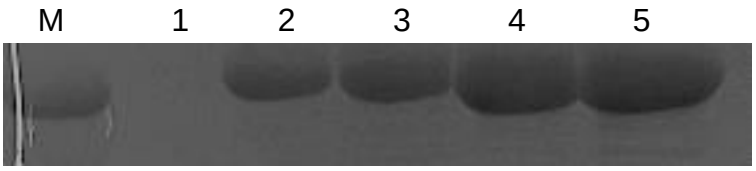


Figura 33. Teste de expressão de Csm2.

SDS-Page de 12% mostrando a banda de 15 kDa do marcador BenchMark™ Protein Ladder (Invitrogen, Carlsbad, EUA) na canaleta M, na canaleta 1 há uma ausência de banda por ser a amostra fermentada sem IPTG, as caneletas subsequentes 2, 3, 4 e 5 mostram as fermentações a 37°C, após 2, 4, 6 e 16 horas de indução com 1 mM IPTG.

4.3 PURIFICAÇÃO DE CSM2 RECOMBINANTE

Para produzir em larga escala a proteína Csm2, o pré-inóculo da bactéria BL21(DE3) foi colocado em quatro litros de meio de cultura LB, mantendo as mesmas condições descritas no teste de expressão (incubação a 37°C, agitação 140 rpm, indução a 1mM IPTG uma vez atingida a densidade ótica a 600nm de 0,5 e coleta de amostra após 6 horas de indução).

O sedimento bacteriano resultante foi lisado com o tampão de lise descrito nos métodos em um homogenizador de alta pressão. A solução foi centrifugada para obter o sobrenadante e purificado por coluna de afinidade de níquel. A proteína posteriormente foi eluída usando um gradiente de tampão B (contendo imidazol). No final do gradiente foi atingida uma concentração de 0,5 M de imidazol correspondendo a 100% do tampão B (figura 34A).

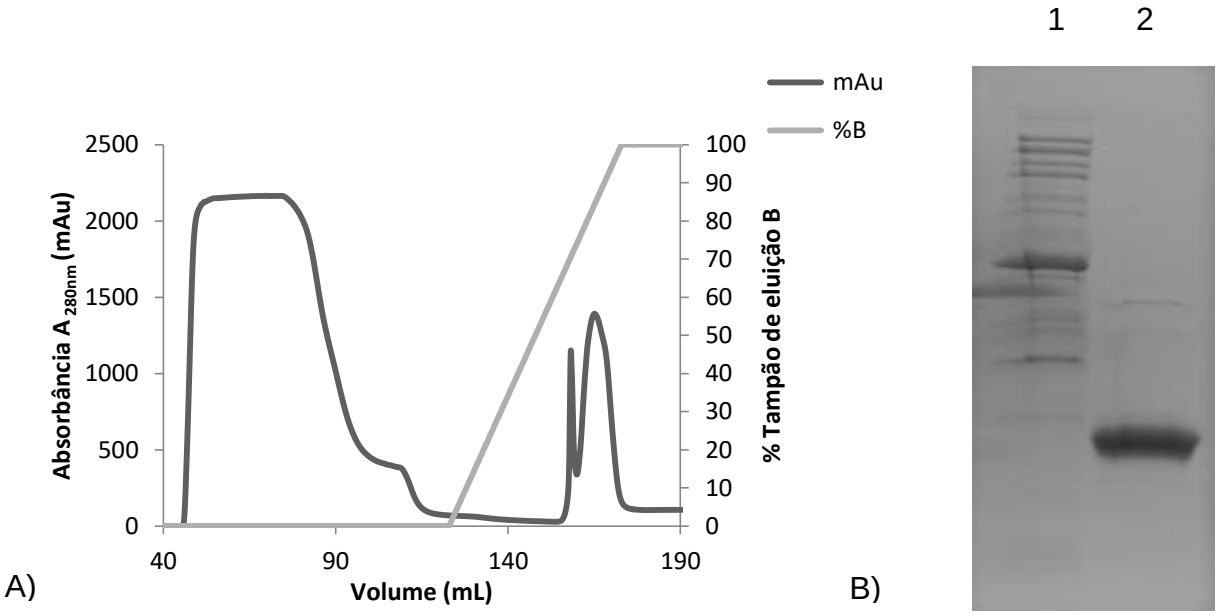


Figura 34. Purificação por afinidade da proteína Csm2.

A) Cromatograma mostrando o pico *flow through* e os picos das proteínas eluídas durante o gradiente de imidazol. B) SDS-Page 12% mostrando a análise dos picos eluídos, primeiro pico com proteínas inespecíficas (canaleta 1) e segundo pico contendo a proteína Csm2 pura (canaleta 2).

Após realizada a purificação, foram coletadas amostras dos dois picos eluídos para ser analisado por eletroforese de proteínas (SDS-PAGE). A figura 34B mostra o padrão de bandas inespecíficas do primeiro pico eluído a 300 mM de imidazol (figura 34B-1) e a proteína eluída em aproximadamente 350 mM de imidazol mostrando uma única banda intensa correspondendo a eluição de Csm2 (figura 34B-2). Esta última fração coletada foi dialisada para retirar o imidazol na amostra.

Esta primeira purificação precisou ser realizada diversas vezes. O motivo foi a brusca precipitação da proteína imediatamente após ser eluída. Já que a proteína era inicialmente eluída em tampão contendo 100 mM NaCl, foi preciso acertar a concentração para 400 mM de NaCl, acrescentando 10% glicerol tanto no tampão de lavagem A como no tampão de eluição B. De esta forma conseguimos evitar a precipitação da Csm2. Tanto o cromatograma como o SDS-PAGE ilustrado são referentes às condições não precipitantes da proteína.

Imediatamente depois da diálise a Csm2 purificada foi submetida a uma clivagem para retirar a sequência de histidinas no extremo N-terminal da proteína. Esta clivagem aconteceu na presença da protease TEV. A separação da proteína de

interesse (Csm2 clivada) das proteínas restantes (proteína Csm2 ainda com a cauda de histidinas e a própria protease TEV) aconteceu realizando uma segunda cromatografia de afinidade, como pode ser evidenciado na figura 35. A proteína clivada carece de histidinas, por tanto ela perde a propriedade de ficar retida na coluna de níquel e forma o primeiro pico do cromatograma, conhecido como *flow through* (figura 35A). Uma amostra referente ao primeiro pico do cromatograma foi analisada por SDS-PAGE, revelando uma massa aproximada de 17 kDa correspondendo a massa esperada (figura 35B).

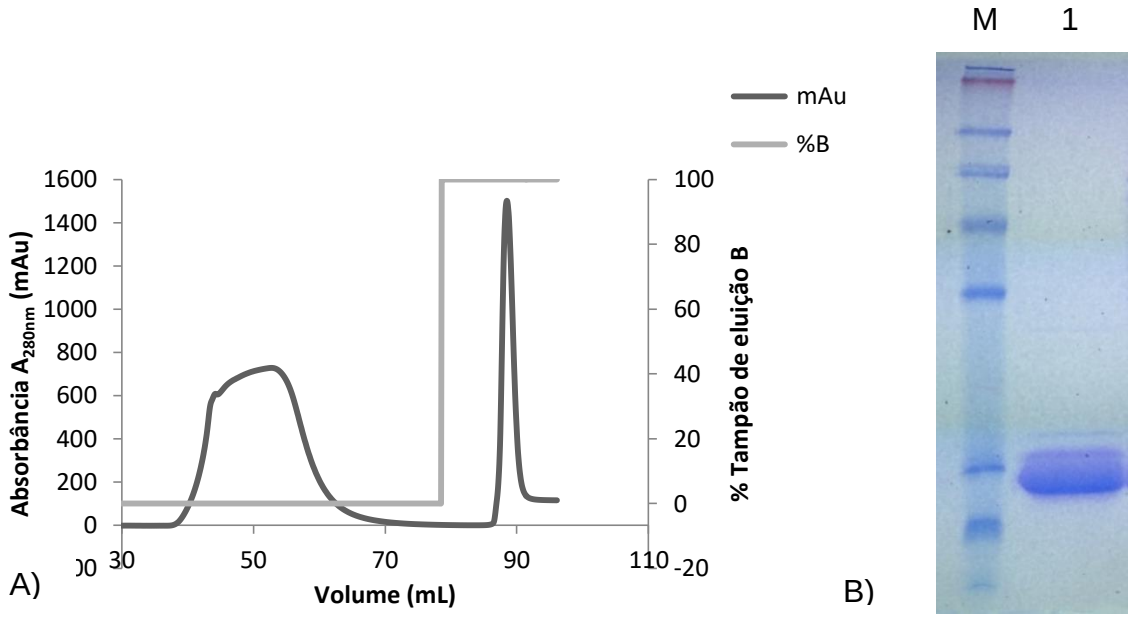


Figura 35. Clivagem de histidinas.

A) Perfil cromatográfico após clivagem de histidinas na porção N-terminal da proteína Csm2. O *flow through* contém a proteína de interesse e o segundo pico eluído de forma isocrática contém restos da proteína Csm2 não clivada e a protease TEV. B) SDS-PAGE de 12% mostrando a análise da proteína clivada (canaleta 1). Marcador SeeBluePlus2 (Invitrogen, Carlsbad, EUA) permite visualizar bandas de 6, 14, 17, 28, 38, 49, 62, 98 e 195 kDa na canaleta M.

A seguir a proteína foi concentrada e aplicada na coluna de gel filtração, desta vez reduzindo a concentração de 400mM NaCl para 100 mM NaCl no tampão A. O cromatograma evidenciou dois picos eluídos os quais foram analisados por eletroforese de proteína. Ambos picos revelaram a mesma massa quando submetidos a condições desnaturantes em um SDS-PAGE (figura 36).

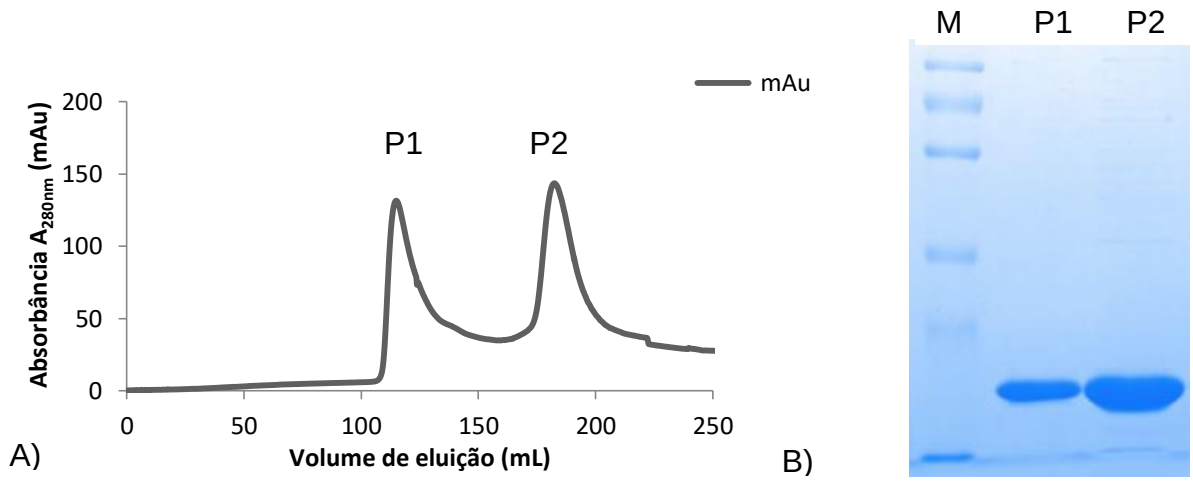


Figura 36. Purificação em gel filtração.

A) Perfil cromatográfico da gel-filtração evidenciando dois picos eluídos. B) SDS-PAGE 12% de ambos picos eluídos mostrando a mesma massa (P1 e P2), o marcador usado foi o LMW Amersham (GE Healthcare, Chicago, EUA) revelando bandas de 15, 20, 30, 45, 66 e 97 kDa (canaleta M).

Ambas frações da proteína eluída foram concentradas até atingir 15 mg/ml. A partir de este momento as duas frações se encontram prontas para realizar os ensaios bioquímicos e cristalográficos pertinentes. Ambas frações foram aliquotadas e armazenadas tanto no *freezer* como no *freezer* -80°C para análises posteriores.

4.4 CRISTALIZAÇÃO DA PROTEÍNA

Para o processo de cristalização foram a utilizados os kits Crystal Screen, Crystal Screen 2 e PEG/Ion (Hampton Research, Aliso Viejo, EUA), sendo que cada solução dos kits está composta por um sal, um tampão e um precipitante diferente. Após realizada a varredura, as placas foram incubadas em estufa à 20°C.

De todas as condições dos kits de cristalização, a condição número 12 do Crystal Screen 2, composta por 100 mM acetato de sódio pH 4,6, 100 mM $CdCl_2$ e 30% PEG 400, apresentou formação de cristais tridimensionais, bem definidos, com bordas regulares após uma semana de incubação como mostrado na figura 37A.

A partir desta condição inicial foram realizados vários refinamentos para melhorar o processo de cristalização. Entre as estratégias utilizadas figuraram aumentar e diminuir a concentração do precipitante PEG 400, trocar o pH do tampão e modificar a concentração do sal $CdCl_2$. Nas figuras 37B, 37C, 37D e 37E podemos observar o melhoramento no processo de refinamento ao diminuir a concentração do precipitante. Entre 21% e 30% do precipitante PEG 400 foram evidenciados

monocristais, porém a 21% de PEG 400 as bordas eram mais definidas e os cristais consideravelmente maiores com menos formação de precipitado na solução da gota (figura 37E).

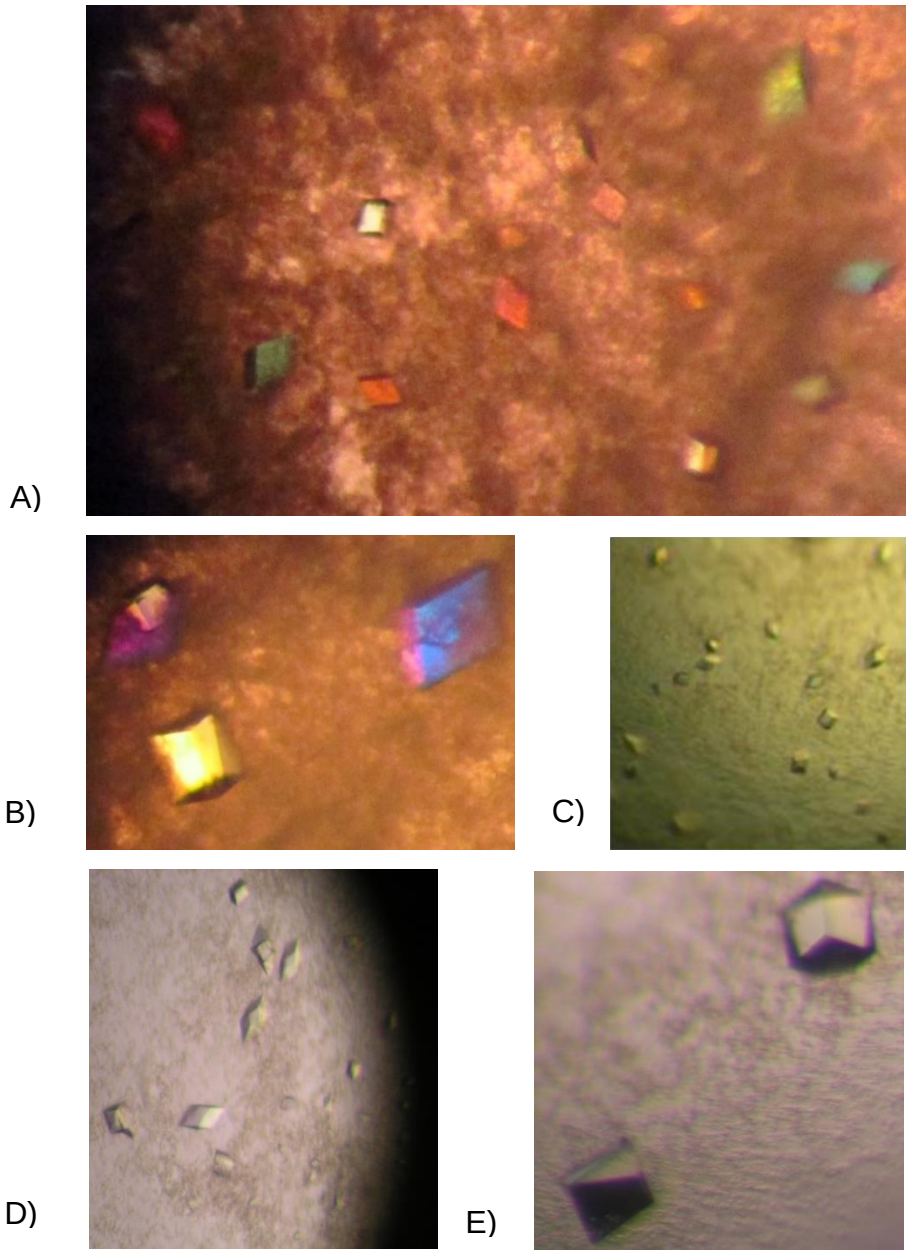


Figura 37. Cristais de proteína Csm2.

A) Cristais de proteína usando o kit Crystal Screen 2 (Hampton Research, Aliso Viejo, EUA) condição número 12. B) 0.1M Cloreto de Cádmio, 0.1M Acetato de Sódio pH 4.6 e 30% PEG400 C) Refinamento usando 25% PEG400. D) Refinamento usando 23% PEG400. E) Refinamento usando 21% PEG400

4.5 COLETA E PROCESSAMENTO DE DADOS CRISTALOGRÁFICOS

Os cristais foram medidos tanto na linha de difração MX-1 como MX-2 do Laboratório Nacional de Luz Síncrotron (LNLS). Segundo os dados obtidos, o cristal nativo (figura 37E) que foi o responsável pela solução da estrutura difratou até 2,9 Å. O padrão de difração do cristal de proteína Csm2 pode ser visualizado na figura 38. A partir dos dados de difração coletados e análise pelo software cristalográfico XDS (Kabsch, 2010), foi possível identificar o grupo espacial do cristal, $P3_121$, com dimensões de célula unitária de a=77 b=77 c=160 e α=90 β=90 γ=120. Conhecendo o peso molecular da proteína, grupo espacial e o volume da célula unitária, é possível realizar o cálculo do coeficiente de Matthews para o cristal usando a seguinte fórmula:

$$\mathrm{Vm} = \frac{\text{Volume de célula unitária}}{\text{Peso molecular da proteína} \times \mathrm{Z} \times \mathrm{X}}$$

Nesta equação o volume de célula unitária é representada em $Å^3$, o peso molecular em é representada em Daltons, Z corresponde ao número de unidades assimétricas na célula unitária e X corresponde ao número de moléculas na unidade assimétrica. Este cálculo sugere cinco soluções (tabela 5), das quais a mais provável sugere a presença de três moléculas de proteína na unidade assimétrica (Vm = 2,7 $Å^3$/Da) com 54% do conteúdo de solvente.

O *Matthews Probability Calculator* pode ser achado no site http://www.ruppweb.org/mattprob/default.html, ele determina o número de moléculas na unidade assimétrica do cristal de proteína baseado no coeficiente de Matthews e comparando no PDB estruturas proteicas a uma resolução menor ou igual a 2,9 Å com as mesmas características cristalográficas da Csm2. Para o cristal de proteína Csm2, de 16,7 kDa, do grupo espacial $P3_121$ que possui 6 unidades assimétricas, é calculado um volume de célula unitária de 821.546,3 $Å^3$. O software indica que o caso mais provável são três moléculas por unidade assimétrica no cristal Csm2 (figura 39).

4.6 DETERMINAÇÃO DA ESTRUTURA

Na linha de difração MX-1 do LNLS foi coletado um conjunto de dados altamente redundantes descritos na tabela 6. Dos cristais difratados, um cristal apresentou um conjunto de dados com uma resolução de 2,9 Å. O grupo espacial pertence ao grupo trigonal $P3_121$.

Como não existem estruturas homólogas a Csm2 descritas no PDB, não foi possível realizar a solução da estrutura por substituição molecular. Não entanto, os cristais de proteínas difratados cresceram em altas concentrações de cloreto de cádmio (100 mM $CdCl_2$). Isto levou a considerar que os íons Cd^{2+} presentes na solução de cristalização possam formar uma ligação com determinados aminoácidos da proteína Csm2, permitindo obtenção de fases usando difração anômala pelo cádmio. É importante ressaltar, que até data, apenas algumas poucas proteínas foram solucionadas usando difração anômala simples (SAD) do cádmio, como a actinidina (Yogavel *et al.*, 2010), a metalotioneína (Robbins *et al.*, 1991) e a ferrochelatase (Medlock *et al.*, 2009).

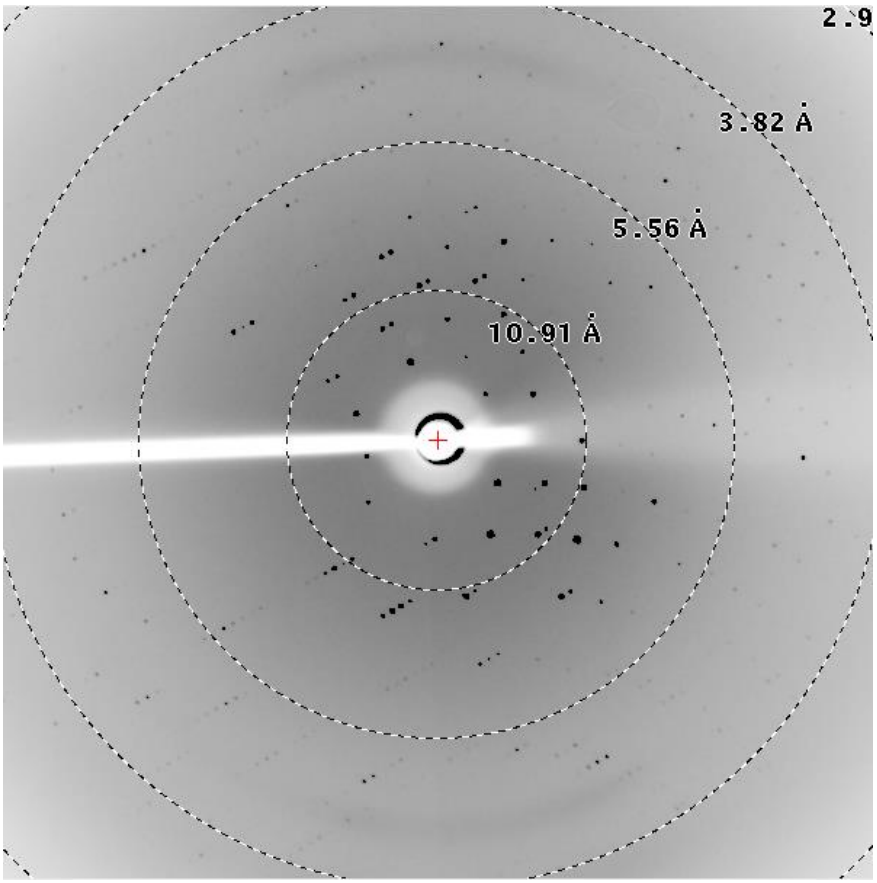


Figura 38. Padrão de difração de raios-X de cristal da proteína Csm2.

Tabela 5. Coeficientes de Matthews para o cristal de Csm2

N (mol)	Vm (A^3/Da)	Vs (% solvente)
1	8.20	85
2	4.10	70
3	**2.73**	**54**
4	2.05	39
5	1.64	24

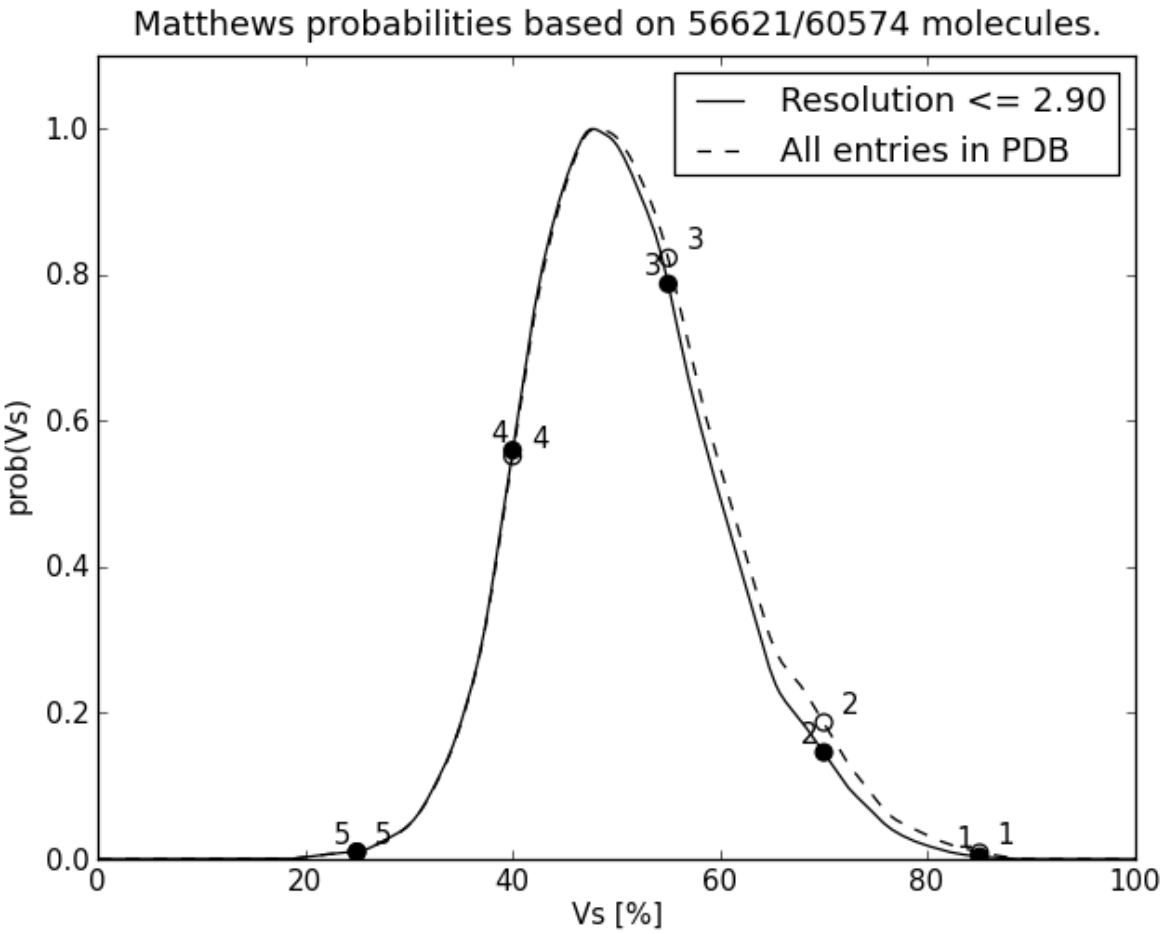


Figura 39. Número de moléculas por unidade assimétrica de Csm2 segundo as probabilidades calculadas do coeficiente de Matthews.

Cálculo realizado online no site http://www.ruppweb.org/mattprob/default.html pelo *Matthews Probability Calculator.*

Tabela 6. Estatística de dados de difração de um cristal de proteína Csm2 na linha MX-1.

Os valores estatísticos para a faixa de mais alta resolução estão entre parênteses.(Gallo *et al.*, 2015)

Fonte de difração	**MX-1, LNLS**
Comprimento de onda (Å)	1,458
Temperatura (K)	100
Detector	MARCCD165
Distância entre cristal e detector (mm)	120
Rotações por imagem (°)	1
Imágens	1-190
Grupo espacial	$P3_121$
Parâmetros de célula unitária *a*, *b*, *c* (Å)	77, 77, 160
Parâmetros de célula unitária α, β, γ (°)	90, 90, 120
Mosaicidade (°)	0,2
Faixa de resolução (Å)	50-2,9 (3,08-2,9)
Número total de reflexões	143 907 (22 874)
Número de reflexões únicos	23 467 (3 813)
Completeza (%)	99,9 (99,6)
Redundância	6,13 (5,99)
$\langle I/\sigma(I)\rangle$	13,77 (2,62)
R_{meas} (%)	11,0 (81,3)
Overall *B* factor from Wilson plot ($Å^2$)	81,2

A difração anômala simples do cádmio (Cd-SAD) foi utilizada em conjunto com o pacote de software cristalográfico PHENIX (Adams *et al.*, 2010) para tentar solucionar a estrutura de Csm2.

De fato, o conjunto de dados medidos e analisados indicaram a presença de sinal anômalo, indicado por uma correlação anômala de em torno de 31% (parâmetro “Anomal Corr” no programa XDS) e uma média de diferenças anômalas

de 1,085 em unidades de desvio padrão (parâmetro “SigAno” no programa XDS), indicando a viabilidade de resolver a estrutura por Cd-SAD.

O módulo phenix.hyss do pacote PHENIX conseguiu detectar 6 íons cádmio com as coordenadas XYZ descritas na tabela 7.

Tabela 7. Coordenadas dos íons cádmio presentes na estrutura cristalográfica da Csm2.

Cádmio	x	y	z
1	29,199	63,826	76,296
2	32,411	57,938	59,546
3	48,581	56,045	68,378
4	21,982	47,716	69,543
5	67,053	48,279	82,014
6	10,816	45,435	36,399

Posteriormente foi realizado um refinamento automático usando o módulo Phaser de PHENIX. Subsequentes modificações de densidade levaram a obtermos o primeiro mapa experimental de densidade eletrônica significativa da estrutura, possibilitando identificar de forma evidente várias α-hélices. Esta solução foi caracterizada por ter dados estatísticos significativos, indicando a solução da estrutura, como um *overall figure of merit* de 0,319 (corresponde ao coseno do erro da fase, onde valores bordeando 0,3 são considerados aceitáveis, valor de 0,4 são considerados bons e valores acima de 0,5 são considerados muito bons), uma correlação de densidade rms local de 0,82 e um *map skew* de 0,10 (valores acima de 0,1 são considerados muito bons para a solução de uma estrutura).

Além de mostrar os dados estatísticos que indicam a solução da estrutura, a própria observação da densidade eletrônica revela a detecção de um padrão α-helical no mapa (figura 40 e 41) a diferentes níveis de contornado e o sinal forte dos átomos cádmio presentes no cristal de proteína.

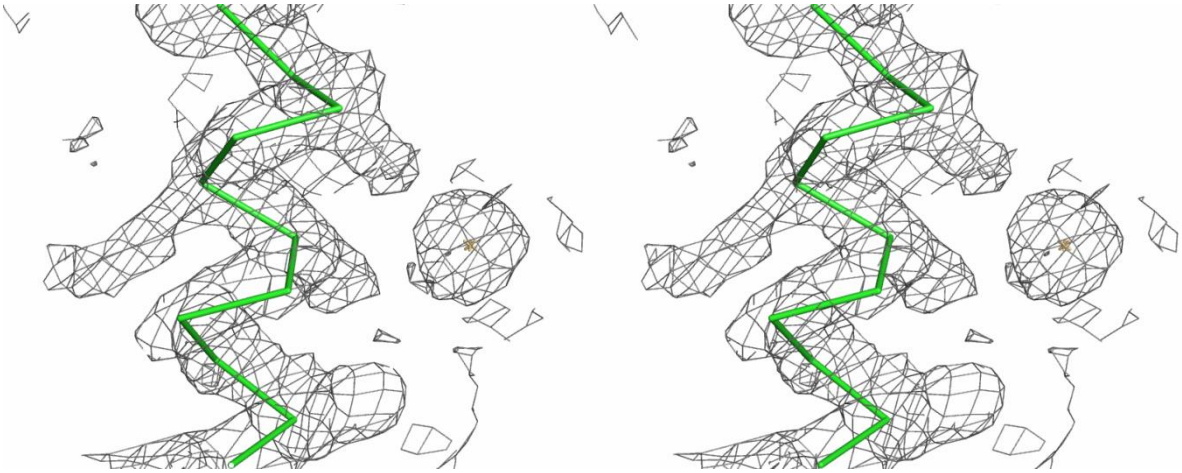


Figura 40. Mapa de densidade eletrônica da estrutura Csm2 um nível de contorno de 1,5σ.

Vista estereográfica mostrando um íon de cádmio (rosa) e a trajetória de uma α hélice (verde).

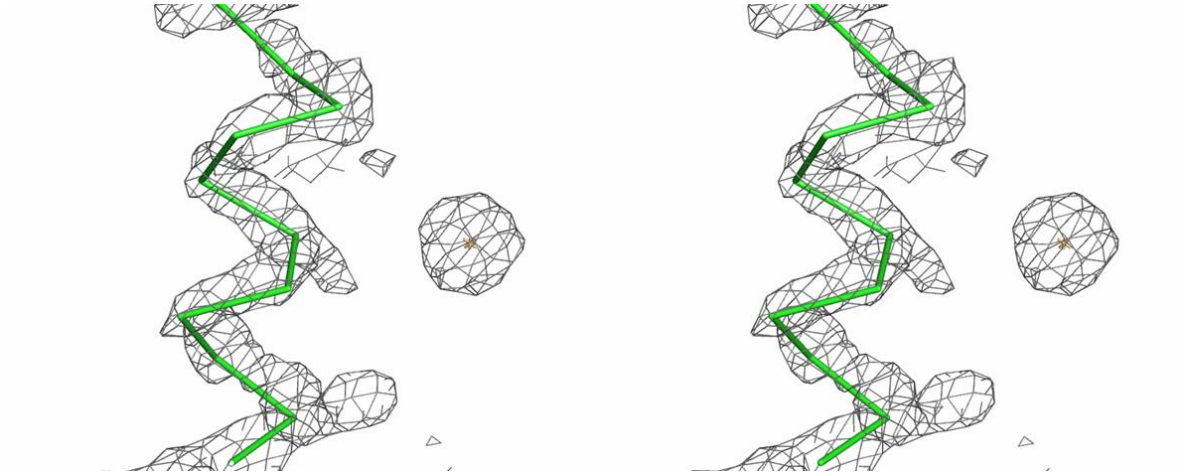


Figura 41. Mapa de densidade eletrônica da estrutura Csm2 um nível de contorno de 3σ.

Vista estereográfica mostrando um íon de cádmio (rosa) e a trajetória de uma α hélice (verde).

4.7 EXTENSÃO DA RESOLUÇÃO

A estrutura com o conjunto de dados coletados a 2,9 Å na linha MX-1 do LNLS permitiu desvendar o mapa de densidade eletrônica, porém a resolução limitada dificultou o assinalamento de cadeias de aminoácidos. Foi preciso coletar outro conjunto de dados cristalográficos, desta vez na linha MX-2 do LNLS para chegar a uma resolução de 2,4 Å que permitiu a elaboração do modelo final da estrutura de Csm2. Este último conjunto de dados está descrito na tabela 8.

Tabela 8. Estatística de dados de difração de um cristal de proteína Csm2 na linha MX-2.

Os valores estatísticos para a faixa de mais alta resolução estão entre parênteses.(Gallo *et al.*, 2016)

Fonte de difração	**MX-2, LNLS**
Comprimento de onda (Å)	1,458
Temperatura (K)	100
Detector	PILATUS2M
Distância entre cristal e detector (mm)	170
Rotações por imagem (°)	1
Imágens	1-126
Grupo espacial	$P3_121$
Parâmetros de célula unitária *a*, *b*, *c* (Å)	77, 77, 160
Parâmetros de célula unitária α, β, γ (°)	90,90, 120
Mosaicidade (°)	0,2
Faixa de resolução (Å)	50-2,4 (2,5-2,4)
Número total de reflexões	135 368 (16 640)
Número de reflexões únicos	21 688 (3 305)
Completeza (%)	98,3 (94,9)
Redundância	6,24 (5,03)
$\langle I/\sigma(I)\rangle$	12,63 (3,64)
R_{meas} (%)	12,9 (83,9)

A tabela 9 mostra uma comparação entre os dois conjuntos de dados em ambas linhas de difração de raios-X do LNLS. Em ambos conjuntos de dados vemos a permanência no grupo espacial $P3_121$, contendo os mesmos parâmetros de célula unitária. Porém no segundo conjunto de dados foi melhorado consideravelmente a resolução, o que permitiu explorar mais o mapa de densidade eletrônica e a construção de aminoácidos dentro do mapa.

Tabela 9. Estatística comparativa entre os dados de difração da solução e refinamento da estrutura da proteína Csm2.

Os valores estatísticos para a faixa de mais alta resolução estão entre parênteses.

Fonte de difração	**MX-1, LNLS**	**MX-2, LNLS**
Comprimento de onda (Å)	1,458	1,458
Temperatura (K)	100	100
Detector	MARCCD165	PILATUS2M
Distância entre cristal e detector (mm)	120	170
Rotações por imagem (°)	1	1
Imágens	1-190	1-126
Grupo espacial	$P3_121$	$P3_121$
Parâmetros de célula unitária *a*, *b*, *c* (Å)	77, 77, 160	77, 77, 160
Parâmetros de célula unitária α, β, γ (°)	90,90, 120	90,90, 120
Mosaicidade (°)	0,2	0,2
Faixa de resolução (Å)	50-2,9 (3,08-2,9)	50-2,4 (2,5-2,4)
Número total de reflexões	143 907 (22 874)	135 368 (16 640)
Número de reflexões únicos	23 467 (3 813)	21 688 (3 305)
Completeza (%)	99,9 (99,6)	98,3 (94,9)
Redundância	6,14 (5,99)	6,24 (5,03)
$\langle I/\sigma(I)\rangle$	13,77 (2,62)	12,63 (3,64)
R_{meas} (%)	11,0 (81,31)	12,9 (83,9)

4.8 REFINAMENTO DO MODELO FINAL

O conjunto de dados cristalográficos obtido na linha de luz MX-2 foi utilizado para o cálculo de mapas de densidade eletrônica de alta resolução, que permitiram a construção e refinamento do modelo final da estrutura. A unidade assimétrica revelou possuir três subunidades de Csm2, como calculado no coeficiente de Matthews. Nas extremidades da molécula não foi possível construir os cinco

primeiros aminoácidos nem os doze últimos aminoácidos da proteína total, isso foi devido à falta de visibilidade dentro do mapa de densidade eletrônica nas extremidades da molécula. O modelo construído possui 123 aminoácidos por cada cadeia, sete íons de cádmio e 112 moléculas de água que foram incorporadas na unidade assimétrica do cristal. O modelo final da estrutura Csm2 foi caraterizado por ter os fatores cristalográficos R*work* e R*free* de 0,2098 e 0,2488, respectivamente, como indicado na tabela 10. Estes valores podem ser considerados como apropriados dada a resolução da estrutura de 2,4 Å.

Tabela 10. Dados estatísticos de refinamento da estrutura Csm2

Os valores estatísticos para a faixa de mais alta resolução estão entre parênteses.

Número de reflexos usados no cálculo de R*work*	21679 (2384)
Número de reflexos usados no cálculo de R*free*	1119 (126)
R_{work} final	0.2098 (0.2352)
R_{free} final	0.2488 (0.2882)
Números de átomos	3248
Número de resíduos de proteína	369
Número de ions (cádmio)	7
Número de moléculas de água	112
RMSD	
Ligações (Å)	0.009
Ângulos (°)	1.111
Média fator *B* (Å^2)	53.10
Diagrama de Ramachandran	
Mais favorecidos (%)	97.25
Permitidos (%)	2.75
Aberrantes (%)	0

4.9 O ENOVELAMENTO DE CSM2

Cada cadeia de Csm2 foi construída entre o aminoácido Gly6 e o aminoácido Ser128. A proteína Csm2 é caracterizada por quatro α-hélices por cadeia (figura 42 e 51). A estrutura da proteína Csm2 de *Thermotoga maritima* foi depositada no PDB com código de acesso 5AN6 (Gallo *et al.*, 2016).

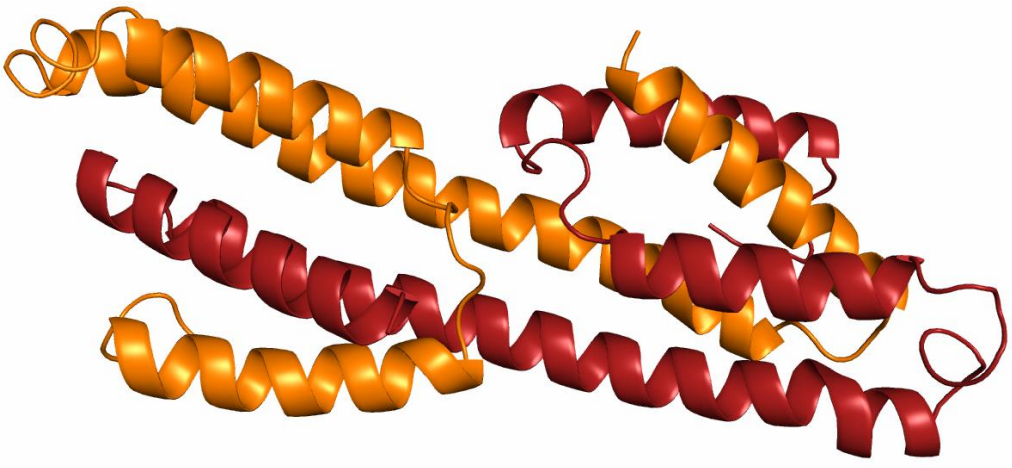


Figura 42. Estrutura do dímero Csm2.

As duas hélices H1 e H2 no extremo N-terminal e a hélice H4 no extremo C-terminal possuem 18, 15 e 18 aminoácidos respectivamente. A α-hélice H3 é a hélice de maior comprimento (possui 42 aminoácidos) (ver figura 43 e 51).

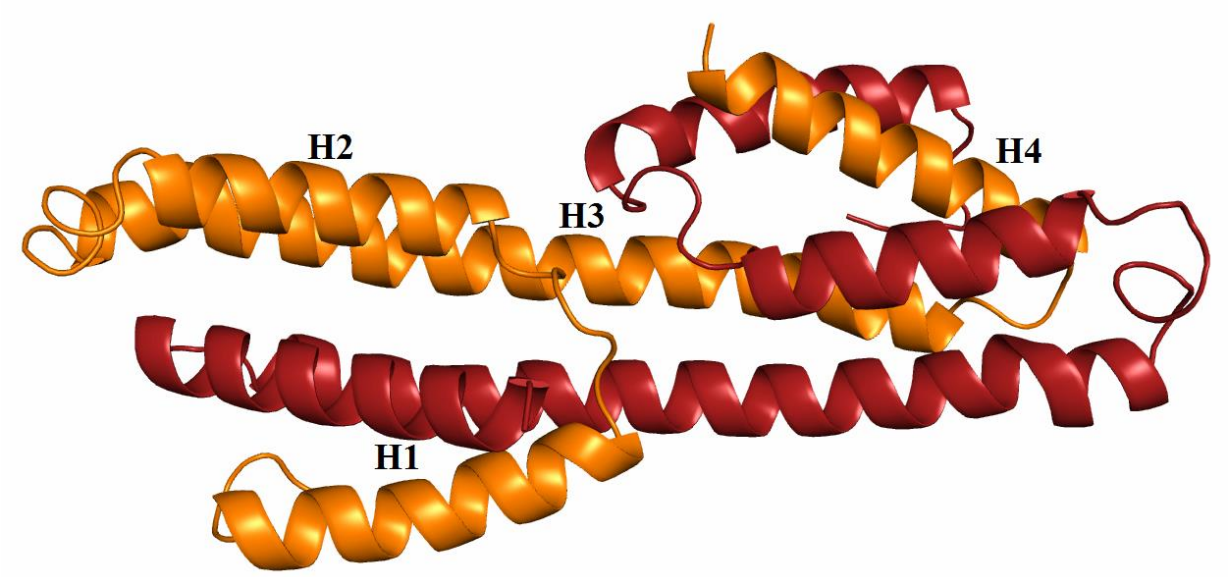


Figura 43. Estrutura do dímero de Csm2 numerando as hélices de uma subunidade.

A hélices como um conjunto formam uma estrutura muito semelhante a um clipe de papel. Nesta estrutura, as α–hélices H1 e H2 se enovelam no N-terminal da hélice maior (H3) e a hélice H4 se enovela no C-terminal da hélice maior. As α–hélices H3 das três cadeias de Csm2 na unidade assimétrica, por sua vez, se encontram alinhadas com outra α–hélice H3 que, no cristal, pode pertencer à mesma unidade assimétrica ou a uma unidade assimétrica adjacente. Este pareamento das α–hélices acontece de maneira antiparalela e induz, portanto, a formação de dímeros por Csm2.

A interação entre duas α–hélices H3 assemelha o visto em estruturas tipo *coiled-coil*, devido à presença de um padrão tipo "*leucine-zipper"* (motivo fecho de leucina) de resíduos hidrofóbicos em posições regulares dentro das hélices. Esta estrutura dimérica é ainda estabilizada por outras interações entre as hélices no N-terminal e C-terminal da proteína. Notavelmente, a α–hélice H1 é inserida numa fenda localizada entre as hélices H3 e H4 da subunidade adjacente. Isto, portanto, leva ao deslocamento da α–hélice H1 de uma subunidade para a outra e a formação de um arranjo de cinco hélices nas extremidades do dímero.

A interação entre as subunidades do dímero é mediado por ligações de hidrogênio e pontes salinas, porém são aparentemente os resíduos hidrofóbicos os principais responsáveis pela forte interação entre estas subunidades. A figura 44 mostra o detalhe de densidade eletrônica da interação de superfície entre as hélices H3 de duas subunidades diferentes. A superfície total calculada da formação da interação do dímero corresponde a 4000 $Å^2$.

4.10 CARACTERIZAÇÃO BIOQUÍMICA

As observações estruturais indicam que a proteína Csm2 forma uma estrutura dimérica como ficou evidente nos cristais de proteína. Esta indicação leva a considerar que o dímero de Csm2 aconteça também *in vitro*. Quando foi realizada a última etapa de purificação por gel filtração, houve uma eluição de duas frações, o que indica a presença de uma forma com peso molecular maior, possivelmente multimérica (primeiro pico) e uma forma com peso molecular menor, possivelmente monomérica (segundo pico, figura 36A). Ambas frações foram coletadas, concentradas e armazenadas. A fração destinada para formação de cristais de

proteína foi o segundo pico, devido a esta última fração corresponder a fração de baixo peso molecular, mesmo que sob condições desnaturantes no gel de SDS-PAGE foi visualizado o mesmo peso molecular para ambas frações (figura 36B).

Para confirmar a possível dimerização da proteína *in vitro,* foram analisadas as frações obtidas tanto por cromatografia de exclusão molecular como por espectrometria de massas. Inicialmente, uma segunda rodada de cromatografia de exclusão molecular confirmou a presença e, portanto, a estabilidade de ambas frações. A figura 45 mostra a injeção do primeiro pico eluído (a fração multimérica), que eluiu novamente no mesmo volume de retenção esperado da fração multimérica (aproximadamente 110 mL após injeção da amostra). A figura 46 mostra a injeção do segundo pico eluído (a fração possivelmente monomérica), que eluiu novamente no mesmo volume de retenção esperado para sua fração monomérica (em torno de 200 mL após injeção de amostra).

A fração escolhida para realizar os cristais de proteína foi a fração de peso molecular menor. Esta fração foi analisada por espectrometria de massas para determinar sua massa molecular. O espectro de massas pode ser visualizado na figura 47. Deconvolução do espectro possibilitou obter uma massa média molecular isotópica de 16.732,78 Da. Este resultado confirma que a fração com peso molecular menor é monomérica.

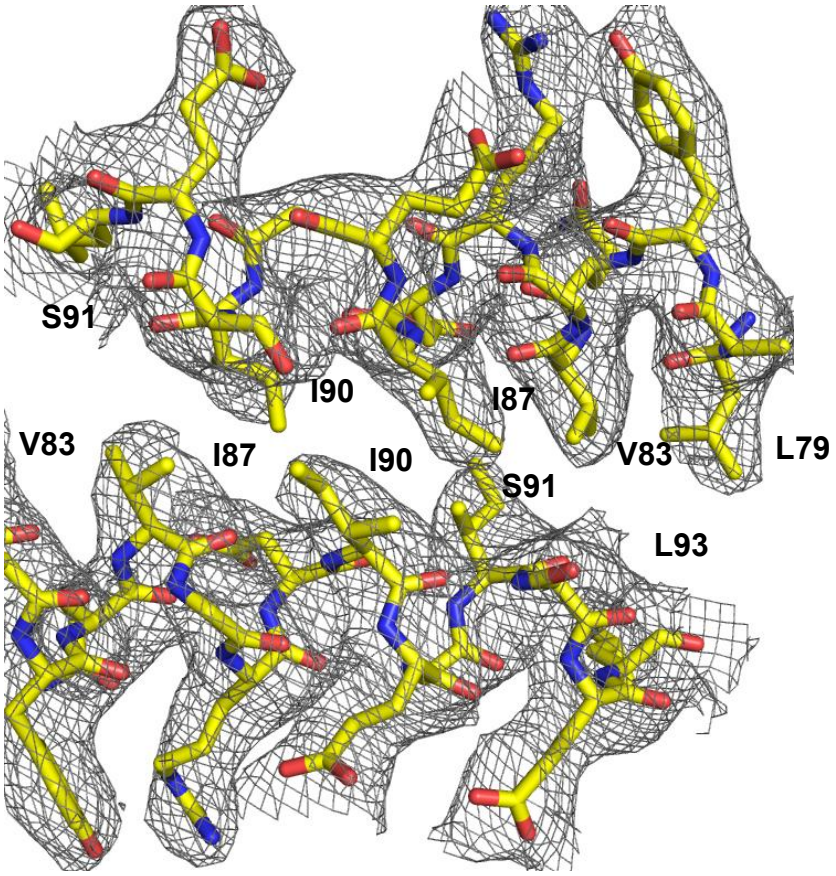


Figura 44. Detalhe do núcleo hidrofóbico do dímero Csm2 mostrando a interação entre as hélices H3.

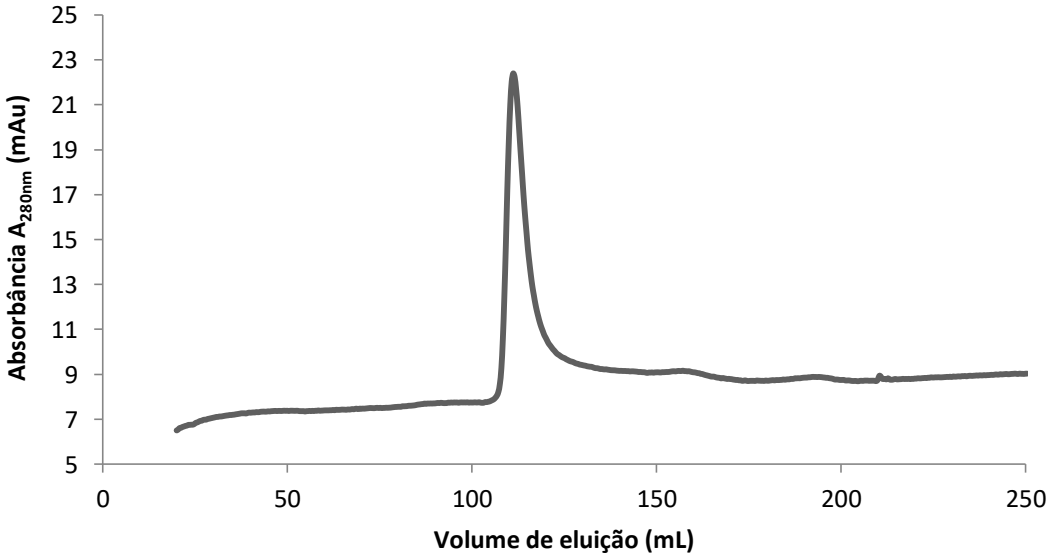


Figura 45. Cromatografia de exclusão molecular da primeira fração demonstrando o multímero intacto.

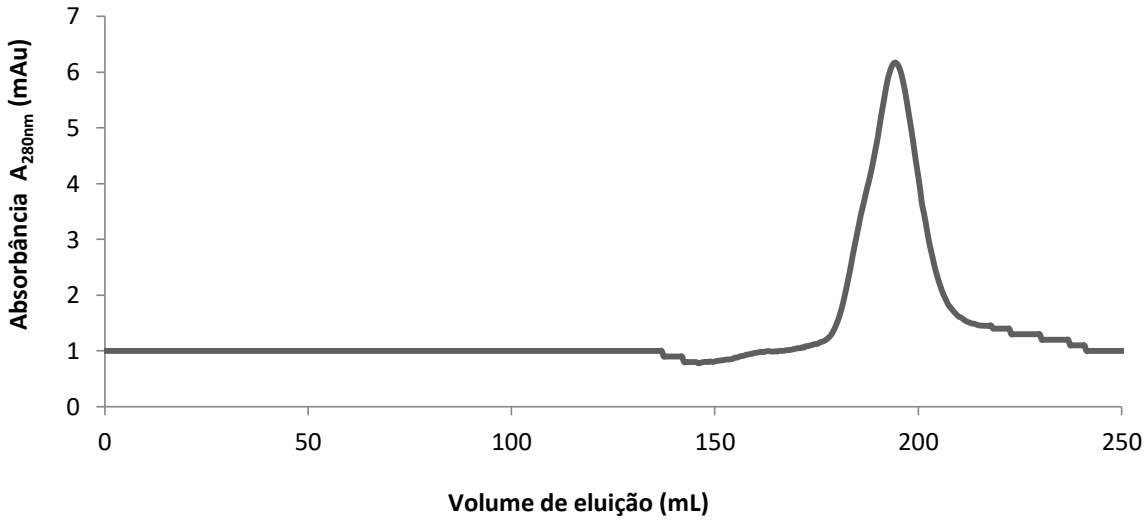


Figura 46. Cromatografia de exclusão molecular da segunda fração demonstrando a presença de monómeros.

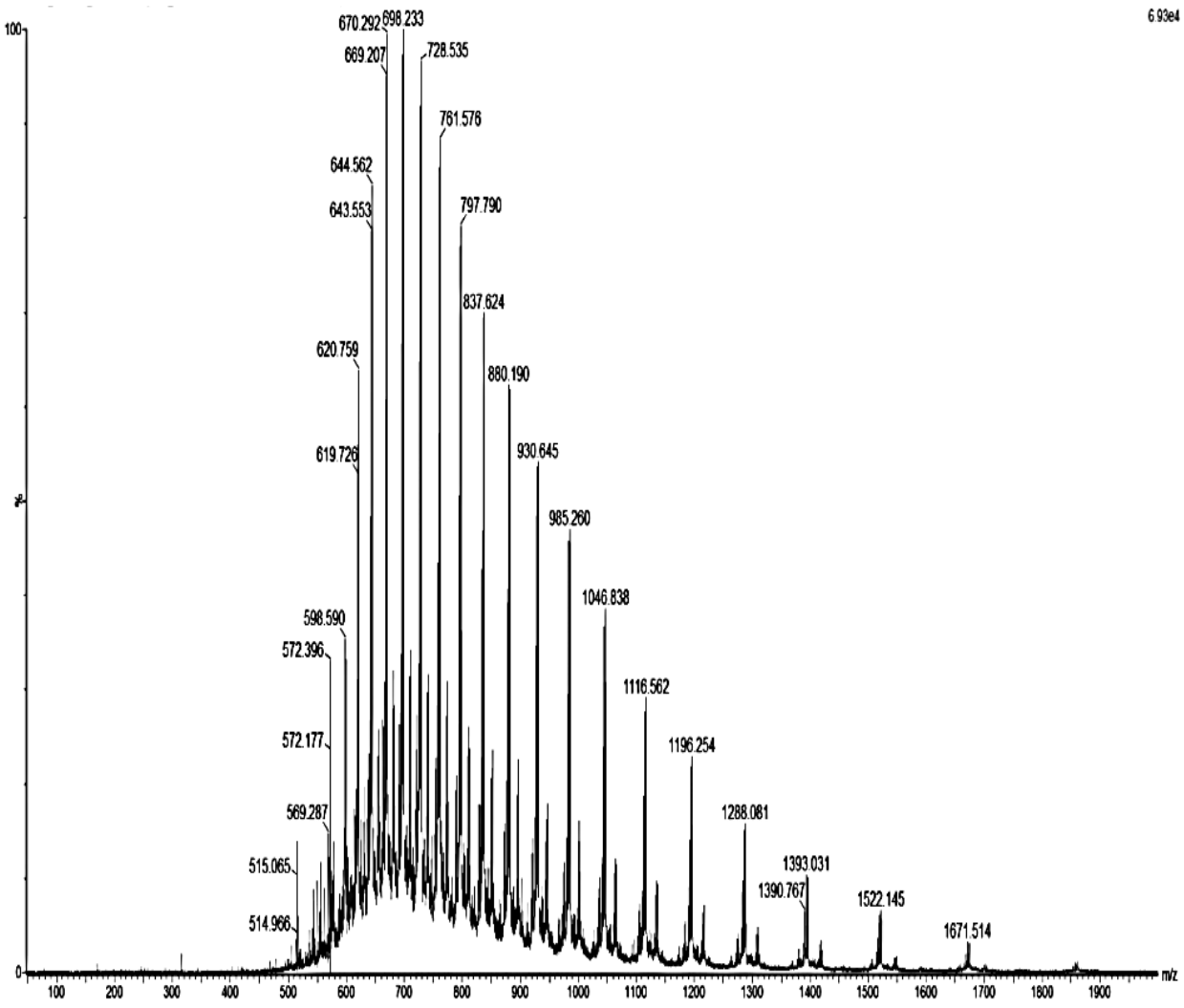


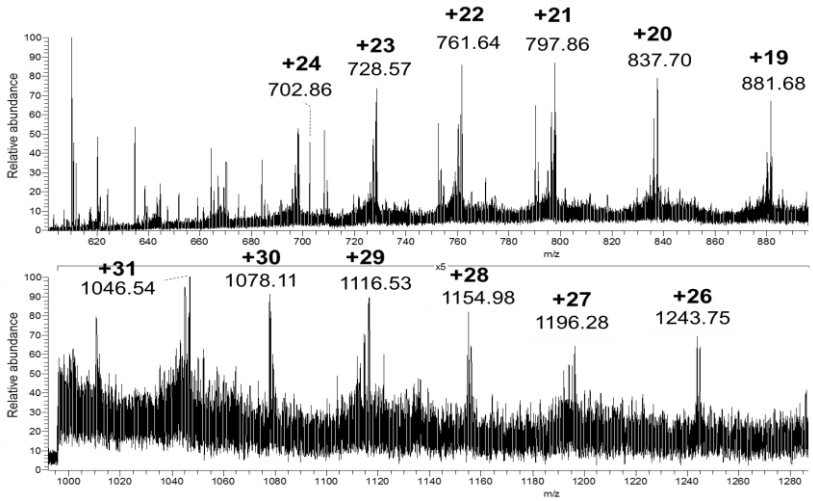
Figura 47. Determinação da massa molecular da proteína nativa Csm2.

A deconvolução do espectro MS gerou uma massa média molecular isotópica de 16.732,78 Da

Posteriormente foi realizado outra análise de espectrometria de massas, desta vez para determinar as massas de ambas frações eluídas por cromatografia de exclusão molecular. A deconvolução do espectro MS da primeira fração, mostrou massas moleculares isotópicas compatíveis tanto para formas monoméricas como para formas diméricas. A distribuição das cargas determinou massas de 16.752,13 ± 4,52 Da e 32.328,25 ± 47,60 Da respectivamente (figura 48). Quando foi analisada a segunda fração por MS, deconvolução do espectro rendeu uma massa molecular isotópica para formas monoméricas com 16.733,10 ± 1,18 Da (figura 49). Os dados obtidos por espectrometria de massas confirmam assim a dimerização da proteína Csm2 nos experimentos de cromatografia líquida.

Massa molecular deconvoluída

Monômero (16.752,13 ± 4,52 Da)



Dímero (32.328,25 ± 47,6 Da)

Figura 48. Análise LC-MS da primeira fração eluída.

A deconvolução permitiu a identificação tanto do estado monomérico como do estado dimérico da proteína Csm2.

Tanto o resultado estrutural, como a caracterização bioquímica demonstram por tanto, a formação de dímeros da proteína Csm2 *in vitro*. As características desta interação, por sua vez, indicam fortemente que o dímero de Csm2 é funcional *in vivo*. Estes dímeros podem contribuir na formação de complexos crRNP Csm e por

consequência na sua função de reconhecimento de oligonucleotídeos alvos para sua posterior degradação.

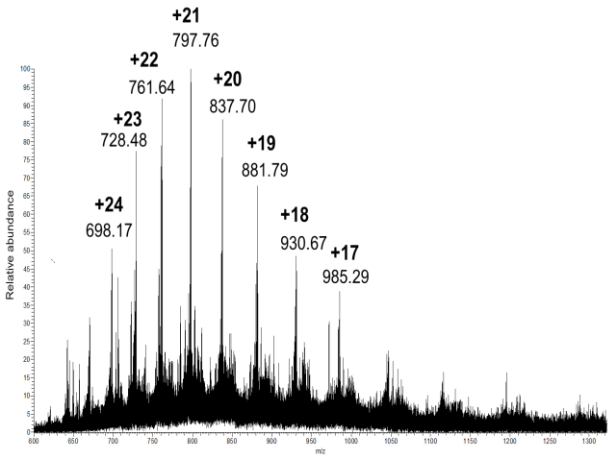


Figura 49. Análise LC-MS da segunda fração eluída.

A deconvolução permitiu a identificação exclusiva do estado monomérico da proteína Csm2.

5 DISCUSSÃO

A solução de estruturas cristalográficas dos complexos crRNP Cascade e Cmr representou um grande avanço para o entendimento do funcionamento dos sistemas CRISPR-Cas dos subtipos I-E e III-B (Staals *et al.*, 2013, Taylor *et al.*, 2015, Jore *et al.*, 2011, Wiedenheft *et al.*, 2011a, Jackson *et al.*, 2014, Zhao *et al.*, 2014, Mulepati *et al.*, 2014, Hayes *et al.*, 2016, Osawa *et al.*, 2015, Spilman *et al.*, 2013, Lintner *et al.*, 2011, Mulepati and Bailey, 2011)**.** Em relação aos complexos crRNA de sistemas CRISPR-Cas tipo III-A, até agora foram solucionadas por cristalografia, apenas proteínas isoladas, como é o caso da proteína Csm1 do organismo *Thermococcus onnurineus* (Jung *et al.*, 2015), a proteína Csm3 do organismo *Methanopyrus kandleri* (Hrle *et al.*, 2013) e a proteína Csm4 de *Methanocaldococcus jannaschii* (Numata *et al.*, 2015). Antes de ser realizado este trabalho não havia estruturas descritas da proteína Csm2 nem da proteína Csm5. Em relação ao complexo Csm, apenas estudos de baixa resolução por microscopia eletrônica, mostrando a composição do complexo Csm de duas espécies procarióticas diferentes, a *S. solfataricus* (Rouillon *et al.*, 2013) e *T. thermophilus* (Staals *et al.*, 2014) foram publicados.

Neste trabalho caracterizamos estruturalmente a proteína Csm2 do sistema CRISPR-Cas subtipo III-A do organismo *Thermotoga maritima* MB8. Como a proteína não demonstra homologia aparente com outras proteínas Cas e não tem homólogos estruturais, a estrutura da proteína foi resolvida por determinação experimental das fases usando o método de Cd-SAD. A proteína tem um enovelamento diferenciado formado por quatro α-hélices. Estas hélices formam dímeros na estrutura. Além da caracterização estrutural desses dímeros, confirmamos a formação deles *in vitro* por estudos de espectrometria de massas e cromatografia liquida de gel-filtração. Estes estudos *in vitro* indicam fortemente que a dimerização de Csm2 também ocorre *in vivo*.

Um mapa de potencial eletroestático mostra o dímero Csm2 possuindo uma face com potencial eletroestático positivo e outra face (realizando uma rotação de 180° no dímero) com cargas de potencial eletroestático negativo (figura 50). Este achado indica que a face carregada com potencial eletroestático positivo (indicado em azul) pode estar envolvida na ligação de oligonucleotídeos.

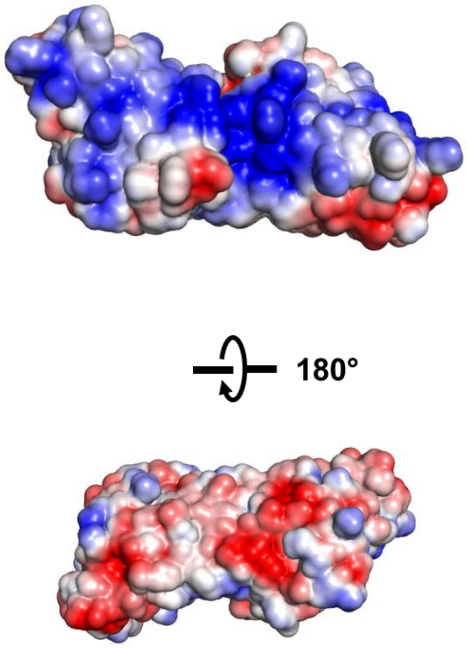


Figura 50. Representação da superfície molecular de potencial eletroestático do dímero Csm2.

Imagem mostrando a mesma orientação representada na figura 42 mostrando um potencial eletroestático positivo (cor azul) e imagem em rotação de 180º referente a ela mesma mostrando um potencial eletroestático negativo (cor vermelha).

A proteína Csm2 de *T. maritima* é um homólogo distante de *S. solfataricus* SSo1424 (18% de identidade na sequência) e *T. thermophilus* Csm2 (22% de identidade). Devido a esta homologia, usamos a estrutura da proteína Csm2 de *T. maritima* para obter mais informações sobre a composição estrutural e função da proteína Csm2 dentro do complexo crRNP subtipo III-A.

Uma análise por CLUSTALΩ mostra que mesmo possuindo uma baixa identidade de aminoácidos global, estas três proteínas de *T. maritima S. solfataricus* e *T. thermophilus* podem ser homólogos funcionais e estruturais. Isto pode ser justificado pelo padrão conservado de importantes resíduos do núcleo hidrofóbico dos agrupamentos de α-hélices do dímero de Csm2. Estes resíduos aparecem sublinhados na figura 51).

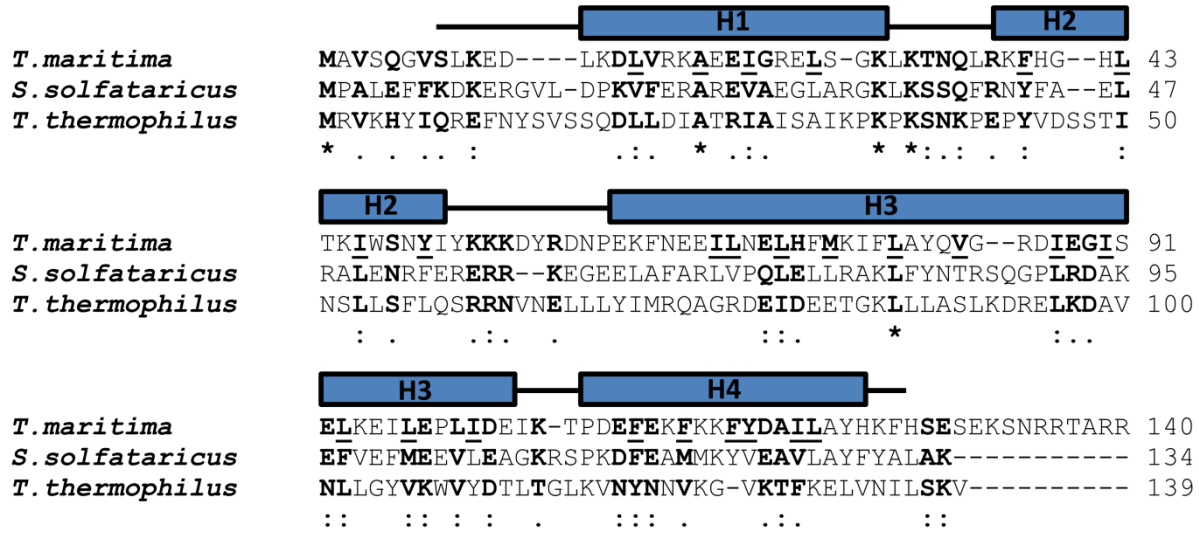


Figura 51. Alinhamento múltiplo de prováveis homólogos de Csm2 a partir de três espécies de procariotos.

As sequências alinhadas mostram uma conservação estrutural entre as proteínas Csm2 de *T. maritima*, SSo1424 de *S. solfataricus* e Cms2 de *T. thermophilus*. Aminoácidos semelhantes estão ressaltados em negrito, o asterisco (*) indica conservação plena, dois pontos (:) indica conservação entre grupos com propriedades fortemente similares e o ponto (.) indica conservação entre grupos com propriedades fracamente similares. Os aminoácidos que formam o núcleo hidrofóbico de Csm2 estão sublinhados.

Nesta análise, as alças que interligam as hélices H1-H2 e H2-H3 mostram um certo grau de conservação, indicando a possibilidade de ter uma importância funcional da molécula.

As estruturas obtidas por microscopia eletrônica dos prováveis homólogos revelam um complexo multimérico de proteínas. Esta estrutura é formada por uma arquitetura entrelaçada helicoidalmente de dois filamentos de proteínas que nas suas extremidades ligam outras proteínas. Estes filamentos são formados pelas subunidades Sso 1424 e Sso1426 (no caso do organismo *S. solfataricus*),e as proteínas Csm2 e Csm3 (em *T. thermophilus*) (Staals *et al.*, 2014, Rouillon *et al.*, 2013). Os complexos resolvidos por microscopia eletrônica do complexo Csm são estruturalmente semelhantes ao complexo Cascade (subtipo I-E) e ao complexo Cmr (Subtipo III-B) que foram resolvidos por cristalografia de proteínas a uma alta resolução (Jackson *et al.*, 2014, Osawa *et al.*, 2015).

Nestas estruturas, a arquitetura é também formada por dois grandes filamentos de proteínas que formam o filamento dorsal e o filamento ventral dos complexos crRNP. O filamento dorsal está envolvido na ligação de crRNA, e o filamento ventral liga os oligonucleotídeos alvos. O filamento dorsal e filamento

ventral são formados principalmente pelas proteínas Cas7 e Cas11, respectivamente, no complexo Cascade (figura 12) (Jackson *et al.*, 2014, Mulepati *et al.*, 2014) e Cmr4 e Cmr5 no complexo Cmr subtipo III-B (figura18) (Osawa *et al.*, 2015). A homologia existente entre a proteína Csm3 e as proteínas Cas7 (do complexo cascade) e Cmr4 (do complexo Cmr) indicam que a proteína Csm3 forma o filamento dorsal do complexo Csm.

Estruturalmente não foi possível confirmar a similaridade entre a proteína Csm2 e as outras proteínas Cas pertencentes ao filamento ventral do complexo crRNP, tal como a Cas11 (subtipo I-E), Cmr5 (subtipo III-B) ou Csa5 (subtipo I-A) (Reeks *et al.*, 2013a). Todas estas proteínas possuem um enovelamento α-helical globular, porém elas não possuem uma α-hélice longa central, e suas hélices adjacentes estão organizadas de forma diferente (figura 52).

Porém, existe a possibilidade que a proteína Csm2 de *T. maritima*, a Sso1424 de *S. solfataricus* e a Csm2 de *T. thermophilus* funcionem de forma análoga a proteína Cas11 de Cascade e Cmr5 do complexo crRNP Cmr na ligação com oligonucleotídeos alvos (Jackson *et al.*, 2014, Mulepati *et al.*, 2014, Osawa *et al.*, 2015)

Neste caso, a similaridade das subunidades homólogas do tipo III-A ao filamento ventral do complexo, tanto do tipo I como do tipo III-B pode ser mais funcional do que estrutural, tal como postulado por Reeks *et al.* (Reeks *et al.*, 2013b).

Para analisar melhor esta hipótese, a estrutura de Csm2 foi incorporada no mapa de densidade eletrônica de *S. solfataricus* (figura 53) (Rouillon *et al.*, 2013) e dentro do mapa de densidade eletrônica de *T. thermophilus* (figura 54) (Staals *et al.*, 2014). Para ambas figuras é possível encaixar dois dímeros de Csm2 dentro da densidade eletrônica, formando um tetrâmero de Csm2 na região que corresponde a ao filamento ventral onde a Sso1424 de *S. solfataricus* ou Csm2 de *T. thermophilus* estariam alocados. (Staals *et al.*, 2014, Rouillon *et al.*, 2013).

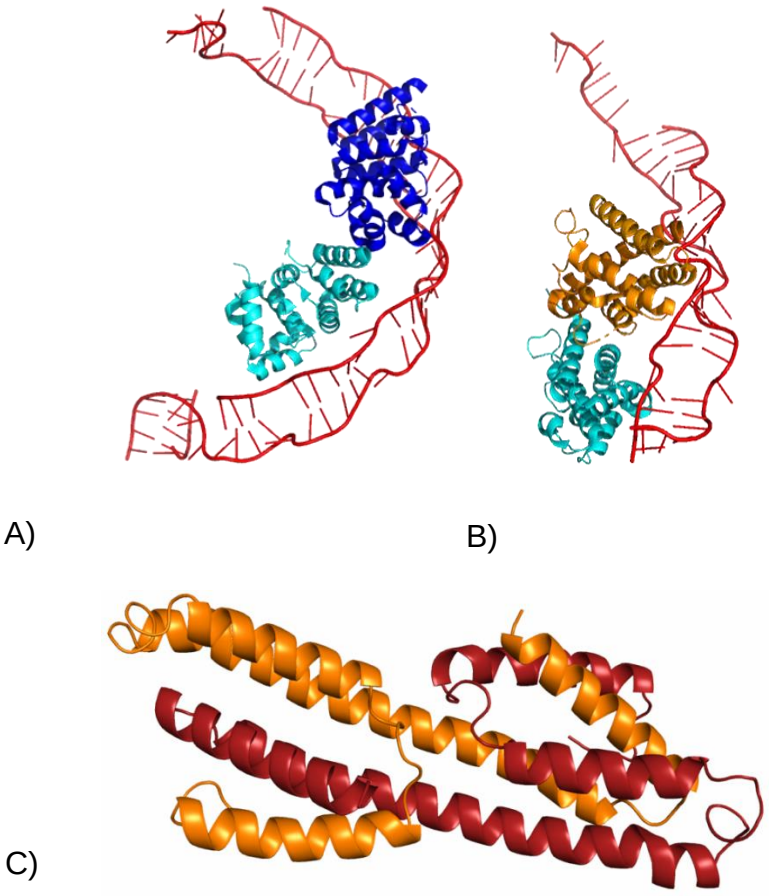


Figura 52. Estruturas análogas formadoras da "barriga" do complexo cnRNP.

A) Dímero Cas11 do complexo Cascade. Organismo *E. coli*.PDB 4QYZ (Mulepati *et al.*, 2014). B) Dímero Cmr5 do complexo Cmr. Organismo *A. fulgidus*. PDB: 3X1L (Osawa *et al.*, 2015). C) Dímero Csm2 que pertence ao complexo Csm. Organismo *T. maritima*. PDB: 5AN6 (Gallo *et al.*, 2016).

Este modelo indica que uma vez mais, que Sso1224 de *S. solfataricus* e Csm2 de *T. thermophilus* podem ser considerados homólogos tanto estruturais como funcionais da proteína Csm2 de *Thermotoga maritima*. Esta informação é consistente com outros resultados demonstrando que o filamento de subunidades pequenas na região ventral pode possivelmente variar no caso do complexo Cmr (Osawa *et al.*, 2015, Spilman *et al.*, 2013) e no caso de complexos Csm. Em ambos complexos o número de subunidades no filamento dorsal pode eventualmente se adaptar para diversos tamanhos de crRNA (Hatoum-Aslan *et al.*, 2013). A formação de dímeros ou tetrâmetros de Csm2, dependendo do comprimento do crRNA, seria compatível com esta hipótese.

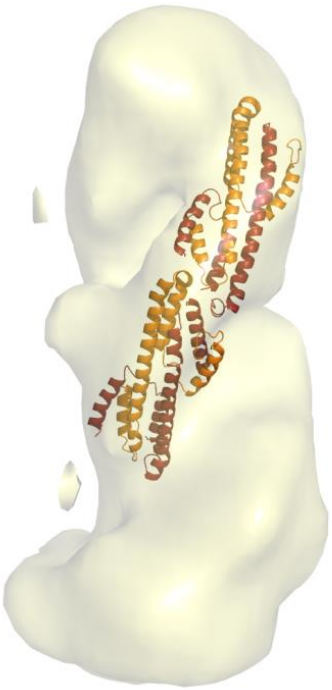


Figura 53. Encaixe da estrutura cristalina de Csm2 dentro do mapa de densidade eletrônica do complexo crRNP de S. solfataricus. EMD-2420 (resolução de 30 Å)

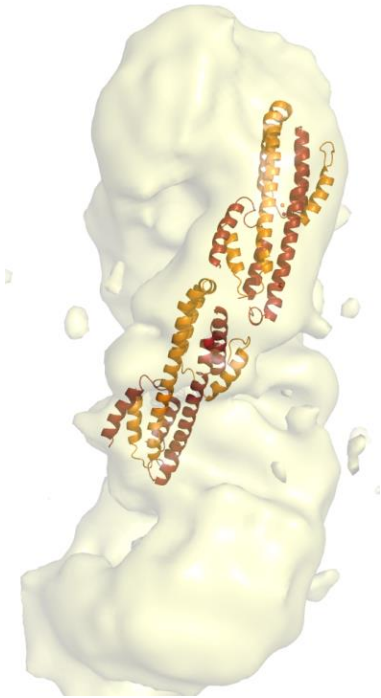


Figura 54. Encaixe da estrutura cristalina de Csm2 dentro do mapa de densidade eletrônica do complexo crRNP de T. thermophilus. EMD-6122, (resolução de 17 Å).

6 CONCLUSÕES

Neste trabalho foi possível resolver a estrutura da proteína Csm2 de *T. maritima* graças a difração anômala simples de cádmio. Esta estrutura atingiu uma resolução de 2,4 Å.

A Csm2 é uma proteína importante dos complexos crRNP do tipo III-A, envolvida muito provavelmente na assemblagem e ligação de oligonucleotídeos alvos do complexo crRNP. Ao contrário da maioria das proteínas participantes de complexos crRNP, sua estrutura não era conhecida, tornando Csm2, a proteína que define complexo IIIA, a última proteína a ser resolvida para compreender o funcionamento desta impressionante maquinaria de RNAi procariótico.

Os resultados obtidos neste trabalho indicam que Csm2 é uma proteína α-helicoidal, com um enovelamento pouco usual e diferente dos enovelamentos conhecidos em proteínas Cas, indicando que possivelmente é um análogo estrutural de outras proteínas em outros complexos e não uma proteína homóloga às mesmas. Mesmo não possuindo homologia entre a proteína Csm2 e as restantes subunidades pequenas dos outros complexos crRNP, os resultados indicam uma analogia funcional entre esta proteína e a Cas11 do complexo Cascade e a Cmr5 do complexo Cmr, indicando que Csm2 pode ser considerado um exemplo de evolução convergente.

Ainda, elucidamos que Csm2 forma dímeros, e conseguimos dissecar as bases estruturais para este fenômeno, que provavelmente é realizado *in vivo* e *in vitro*. A dimerização da Csm2 descrita aqui é importante para entender a função dos complexos crRNP subtipo III-A do sistema CRISPR-Cas.

7 REFERÊNCIAS

ABERGEL, C. 2013. Molecular replacement: tricks and treats. *Acta Crystallogr D Biol Crystallogr,* 69**,** 2167-73.

ABUDAYYEH, O. O., GOOTENBERG, J. S., KONERMANN, S., JOUNG, J., SLAYMAKER, I. M., COX, D. B., SHMAKOV, S., MAKAROVA, K. S., SEMENOVA, E., MINAKHIN, L., SEVERINOV, K., REGEV, A., LANDER, E. S., KOONIN, E. V. & ZHANG, F. 2016. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science,* 353**,** aaf5573.

ADAMS, P. D., AFONINE, P. V., BUNKOCZI, G., CHEN, V. B., DAVIS, I. W., ECHOLS, N., HEADD, J. J., HUNG, L. W., KAPRAL, G. J., GROSSE-KUNSTLEVE, R. W., MCCOY, A. J., MORIARTY, N. W., OEFFNER, R., READ, R. J., RICHARDSON, D. C., RICHARDSON, J. S., TERWILLIGER, T. C. & ZWART, P. H. 2010. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr,* 66**,** 213-21.

AFONINE, P. V., GROSSE-KUNSTLEVE, R. W., ECHOLS, N., HEADD, J. J., MORIARTY, N. W., MUSTYAKIMOV, M., TERWILLIGER, T. C., URZHUMTSEV, A., ZWART, P. H. & ADAMS, P. D. 2012. Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr D Biol Crystallogr,* 68**,** 352-67.

ANDERS, C., NIEWOEHNER, O., DUERST, A. & JINEK, M. 2014. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature,* 513**,** 569-73.

ARVAI, A. 2015. *Adxv: A program to display X-Ray diffraction images* [Online]. Available: http://www.scripps.edu/tainer/arvai/adxv.html [Accessed].

BAKER, N. A., SEPT, D., JOSEPH, S., HOLST, M. J. & MCCAMMON, J. A. 2001. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A,* 98**,** 10037-41.

BANEYX, F. 1999. Recombinant protein expression in Escherichia coli. *Curr Opin Biotechnol,* 10**,** 411-21.

BARRANGOU, R., FREMAUX, C., DEVEAU, H., RICHARDS, M., BOYAVAL, P., MOINEAU, S., ROMERO, D. A. & HORVATH, P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science,* 315**,** 1709-12.

BARRANGOU, R. & HORVATH, P. 2012. CRISPR: new horizons in phage resistance and strain identification. *Annu Rev Food Sci Technol,* 3**,** 143-62.

BELL, M. R., ENGLEKA, M. J., MALIK, A. & STRICKLER, J. E. 2013. To fuse or not to fuse: what is your purpose? *Protein Sci,* 22**,** 1466-77.

BOHACIAKOVA, D., RENZOVA, T., FEDOROVA, V., BARAK, M., KUNOVA BOSAKOVA, M., HAMPL, A. & CAJANEK, L. 2017. An Efficient Method for Generation of Knockout Human Embryonic Stem Cells Using CRISPR/Cas9 System. *Stem Cells Dev,* 26**,** 1521-1527.

BOLOTIN, A., QUINQUIS, B., SOROKIN, A. & EHRLICH, S. D. 2005. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology,* 151**,** 2551-61.

BRINER, A. E., DONOHOUE, P. D., GOMAA, A. A., SELLE, K., SLORACH, E. M., NYE, C. H., HAURWITZ, R. E., BEISEL, C. L., MAY, A. P. & BARRANGOU, R. 2014. Guide RNA functional modules direct Cas9 activity and orthogonality. *Mol Cell,* 56**,** 333-9.

BROUNS, S. J., JORE, M. M., LUNDGREN, M., WESTRA, E. R., SLIJKHUIS, R. J., SNIJDERS, A. P., DICKMAN, M. J., MAKAROVA, K. S., KOONIN, E. V. & VAN DER OOST, J. 2008. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science,* 321**,** 960-4.

BRUNGER, A. T. 2007. Version 1.2 of the Crystallography and NMR system. *Nat Protoc,* 2**,** 2728-33.

BRUNGER, A. T., ADAMS, P. D., CLORE, G. M., DELANO, W. L., GROS, P., GROSSE-KUNSTLEVE, R. W., JIANG, J. S., KUSZEWSKI, J., NILGES, M., PANNU, N. S., READ, R. J., RICE, L. M., SIMONSON, T. & WARREN, G. L. 1998. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr,* 54**,** 905-21.

CASS, S. D., HAAS, K. A., STOLL, B., ALKHNBASHI, O. S., SHARMA, K., URLAUB, H., BACKOFEN, R., MARCHFELDER, A. & BOLT, E. L. 2015. The role of Cas8 in type I CRISPR interference. *Biosci Rep,* 35.

CHYLINSKI, K., LE RHUN, A. & CHARPENTIER, E. 2013. The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. *RNA Biol,* 10**,** 726-37.

CHYLINSKI, K., MAKAROVA, K. S., CHARPENTIER, E. & KOONIN, E. V. 2014. Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Res,* 42**,** 6091-105.

CONG, L., RAN, F. A., COX, D., LIN, S., BARRETTO, R., HABIB, N., HSU, P. D., WU, X., JIANG, W., MARRAFFINI, L. A. & ZHANG, F. 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science,* 339**,** 819-23.

CREGG, J. M., CEREGHINO, J. L., SHI, J. & HIGGINS, D. R. 2000. Recombinant protein expression in Pichia pastoris. *Mol Biotechnol,* 16**,** 23-52.

CRESS, B. F., TOPARLAK, O. D., GULERIA, S., LEBOVICH, M., STIEGLITZ, J. T., ENGLAENDER, J. A., JONES, J. A., LINHARDT, R. J. & KOFFAS, M. A. 2015. CRISPathBrick: Modular Combinatorial Assembly of Type II-A CRISPR Arrays for dCas9-Mediated Multiplex Transcriptional Repression in E. coli. *ACS Synth Biol,* 4**,** 987-1000.

DELANO, W. L. 2015. *The PyMOL Molecular Graphics System* [Online]. Available: https://www.pymol.org/ [Accessed].

DELTCHEVA, E., CHYLINSKI, K., SHARMA, C. M., GONZALES, K., CHAO, Y., PIRZADA, Z. A., ECKERT, M. R., VOGEL, J. & CHARPENTIER, E. 2011. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature,* 471**,** 602-7.

DENG, L., GARRETT, R. A., SHAH, S. A., PENG, X. & SHE, Q. 2013. A novel interference mechanism by a type IIIB CRISPR-Cmr module in Sulfolobus. *Mol Microbiol,* 87**,** 1088-99.

DONG, C., FONTANA, J., PATEL, A., CAROTHERS, J. M. & ZALATAN, J. G. 2018. Synthetic CRISPR-Cas gene activators for transcriptional reprogramming in bacteria. *Nat Commun,* 9**,** 2489.

DRENTH, J. 2007. *Principles of Protein X-Ray Crystallography*.

EMSLEY, P., LOHKAMP, B., SCOTT, W. G. & COWTAN, K. 2010. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr,* 66**,** 486-501.

EVANS, L., HUGHES, M., WATERS, J., CAMERON, J., DODSWORTH, N., TOOTH, D., GREENFIELD, A. & SLEEP, D. 2010. The production, characterisation and enhanced pharmacokinetics of scFv-albumin fusions expressed in Saccharomyces cerevisiae. *Protein Expr Purif,* 73**,** 113-24.

FANG, L., JIA, K. Z., TANG, Y. L., MA, D. Y., YU, M. & HUA, Z. C. 2007. An improved strategy for high-level production of TEV protease in Escherichia coli and its purification and characterization. *Protein Expr Purif,* 51**,** 102-9.

GALLO, G., AUGUSTO, G., RANGEL, G., ZELANIS, A., MORI, M. A., BARBOSA CAMPOS, C. & WURTELE, M. 2015. Purification, crystallization, crystallographic analysis and phasing of the CRISPR-associated protein Csm2 from Thermotoga maritima. *Acta Crystallogr F Struct Biol Commun,* 71**,** 1223-7.

GALLO, G., AUGUSTO, G., RANGEL, G., ZELANIS, A., MORI, M. A., CAMPOS, C. B. & WURTELE, M. 2016. Structural basis for dimer formation of the CRISPR-associated protein Csm2 of Thermotoga maritima. *FEBS J,* 283**,** 694-703.

GARNEAU, J. E., DUPUIS, M. E., VILLION, M., ROMERO, D. A., BARRANGOU, R., BOYAVAL, P., FREMAUX, C., HORVATH, P., MAGADAN, A. H. & MOINEAU, S. 2010. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature,* 468**,** 67-71.

GASIUNAS, G., BARRANGOU, R., HORVATH, P. & SIKSNYS, V. 2012. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc Natl Acad Sci U S A,* 109**,** E2579-86.

GOH, S. & GOOD, L. 2008. Plasmid selection in Escherichia coli using an endogenous essential gene marker. *BMC Biotechnol,* 8**,** 61.

GOLDBERG, G. W., JIANG, W., BIKARD, D. & MARRAFFINI, L. A. 2014. Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. *Nature,* 514**,** 633-7.

GONG, B., SHIN, M., SUN, J., JUNG, C. H., BOLT, E. L., VAN DER OOST, J. & KIM, J. S. 2014. Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3. *Proc Natl Acad Sci U S A,* 111**,** 16359-64.

GRASLUND, S., NORDLUND, P., WEIGELT, J., HALLBERG, B. M., BRAY, J., GILEADI, O., KNAPP, S., OPPERMANN, U., ARROWSMITH, C., HUI, R., MING, J., DHE-PAGANON, S., PARK, H. W., SAVCHENKO, A., YEE, A., EDWARDS, A., VINCENTELLI, R., CAMBILLAU, C., KIM, R., KIM, S. H., RAO, Z., SHI, Y., TERWILLIGER, T. C., KIM, C. Y., HUNG, L. W., WALDO, G. S., PELEG, Y., ALBECK, S., UNGER, T., DYM, O., PRILUSKY, J., SUSSMAN, J. L., STEVENS, R. C., LESLEY, S. A., WILSON, I. A., JOACHIMIAK, A., COLLART, F., DEMENTIEVA, I., DONNELLY, M. I., ESCHENFELDT, W. H., KIM, Y., STOLS, L., WU, R., ZHOU, M., BURLEY, S. K., EMTAGE, J. S., SAUDER, J. M., THOMPSON, D., BAIN, K., LUZ, J., GHEYI, T., ZHANG, F., ATWELL, S., ALMO, S. C., BONANNO, J. B., FISER, A., SWAMINATHAN, S., STUDIER, F. W., CHANCE, M. R., SALI, A., ACTON, T. B., XIAO, R., ZHAO, L., MA, L. C., HUNT, J. F., TONG, L., CUNNINGHAM, K., INOUYE, M., ANDERSON, S., JANJUA, H., SHASTRY, R., HO, C. K., WANG, D., WANG, H., JIANG, M., MONTELIONE, G. T., STUART, D. I., OWENS, R. J., DAENKE, S., SCHUTZ, A., HEINEMANN, U., YOKOYAMA, S., BUSSOW, K. & GUNSALUS, K. C. 2008. Protein production and purification. *Nat Methods,* 5**,** 135-46.

GROENEN, P. M., BUNSCHOTEN, A. E., VAN SOOLINGEN, D. & VAN EMBDEN, J. D. 1993. Nature of DNA polymorphism in the direct repeat cluster of Mycobacterium tuberculosis; application for strain differentiation by a novel typing method. *Mol Microbiol,* 10**,** 1057-65.

GUIMARAES, B. G., SANFELICI, L., NEUENSCHWANDER, R. T., RODRIGUES, F., GRIZOLLI, W. C., RAULIK, M. A., PITON, J. R., MEYER, B. C., NASCIMENTO, A. S. & POLIKARPOV, I. 2009. The MX2 macromolecular crystallography beamline: a wiggler X-ray source at the LNLS. *J Synchrotron Radiat,* 16**,** 69-75.

HAFT, D. H., SELENGUT, J., MONGODIN, E. F. & NELSON, K. E. 2005. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol,* 1**,** e60.

HALE, C. R., MAJUMDAR, S., ELMORE, J., PFISTER, N., COMPTON, M., OLSON, S., RESCH, A. M., GLOVER, C. V., 3RD, GRAVELEY, B. R., TERNS, R. M. & TERNS, M. P. 2012. Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Mol Cell,* 45**,** 292-302.

HALE, C. R., ZHAO, P., OLSON, S., DUFF, M. O., GRAVELEY, B. R., WELLS, L., TERNS, R. M. & TERNS, M. P. 2009. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell,* 139**,** 945-56.

HATOUM-ASLAN, A., MANIV, I. & MARRAFFINI, L. A. 2011. Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is

measured by a ruler mechanism anchored at the precursor processing site. *Proc Natl Acad Sci U S A,* 108**,** 21218-22.

HATOUM-ASLAN, A., MANIV, I., SAMAI, P. & MARRAFFINI, L. A. 2014. Genetic characterization of antiplasmid immunity through a type III-A CRISPR-Cas system. *J Bacteriol,* 196**,** 310-7.

HATOUM-ASLAN, A., SAMAI, P., MANIV, I., JIANG, W. & MARRAFFINI, L. A. 2013. A ruler protein in a complex for antiviral defense determines the length of small interfering CRISPR RNAs. *J Biol Chem,* 288**,** 27888-97.

HAYES, R. P., XIAO, Y., DING, F., VAN ERP, P. B., RAJASHANKAR, K., BAILEY, S., WIEDENHEFT, B. & KE, A. 2016. Structural basis for promiscuous PAM recognition in type I-E Cascade from E. coli. *Nature,* 530**,** 499-503.

HELER, R., SAMAI, P., MODELL, J. W., WEINER, C., GOLDBERG, G. W., BIKARD, D. & MARRAFFINI, L. A. 2015. Cas9 specifies functional viral targets during CRISPR-Cas adaptation. *Nature,* 519**,** 199-202.

HOCHSTRASSER, M. L. & DOUDNA, J. A. 2015. Cutting it close: CRISPR-associated endoribonuclease structure and function. *Trends Biochem Sci,* 40**,** 58-66.

HRLE, A., SU, A. A., EBERT, J., BENDA, C., RANDAU, L. & CONTI, E. 2013. Structure and RNA-binding properties of the type III-A CRISPR-associated protein Csm3. *RNA Biol,* 10**,** 1670-8.

HUBER, R. L., THOMAS A.; KOENIG, HELMUT; THOMM, MICHAEL; WOESE, CARL R.; SLEYTR, UWE B.; STETTER, KARL O. 1986. Thermotoga maritima sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90°C. *Archives of Microbiology,* 144**,** 324-33.

HUO, Y., NAM, K. H., DING, F., LEE, H., WU, L., XIAO, Y., FARCHIONE, M. D., JR., ZHOU, S., RAJASHANKAR, K., KURINOV, I., ZHANG, R. & KE, A. 2014. Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nat Struct Mol Biol,* 21**,** 771-7.

JACKSON, R. N., GOLDEN, S. M., VAN ERP, P. B., CARTER, J., WESTRA, E. R., BROUNS, S. J., VAN DER OOST, J., TERWILLIGER, T. C., READ, R. J. & WIEDENHEFT, B. 2014. Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. *Science,* 345**,** 1473-9.

JANSEN, R., EMBDEN, J. D., GAASTRA, W. & SCHOULS, L. M. 2002. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol,* 43**,** 1565-75.

JIANG, F. & DOUDNA, J. A. 2017. CRISPR-Cas9 Structures and Mechanisms. *Annu Rev Biophys,* 46**,** 505-529.

JINEK, M., CHYLINSKI, K., FONFARA, I., HAUER, M., DOUDNA, J. A. & CHARPENTIER, E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science,* 337**,** 816-21.

JINEK, M., JIANG, F., TAYLOR, D. W., STERNBERG, S. H., KAYA, E., MA, E., ANDERS, C., HAUER, M., ZHOU, K., LIN, S., KAPLAN, M., IAVARONE, A. T., CHARPENTIER, E., NOGALES, E. & DOUDNA, J. A. 2014. Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science,* 343**,** 1247997.

JOHNSON, I. S. 1983. Human insulin from recombinant DNA technology. *Science,* 219**,** 632-7.

JORE, M. M., LUNDGREN, M., VAN DUIJN, E., BULTEMA, J. B., WESTRA, E. R., WAGHMARE, S. P., WIEDENHEFT, B., PUL, U., WURM, R., WAGNER, R., BEIJER, M. R., BARENDREGT, A., ZHOU, K., SNIJDERS, A. P., DICKMAN, M. J., DOUDNA, J. A., BOEKEMA, E. J., HECK, A. J., VAN DER OOST, J. & BROUNS, S. J. 2011. Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat Struct Mol Biol,* 18**,** 529-36.

JUNG, T. Y., AN, Y., PARK, K. H., LEE, M. H., OH, B. H. & WOO, E. 2015. Crystal structure of the Csm1 subunit of the Csm complex and its single-stranded DNA-specific nuclease activity. *Structure,* 23**,** 782-90.

KABSCH, W. 2010. Xds. *Acta Crystallogr D Biol Crystallogr,* 66**,** 125-32.

KENDREW, J. C., BODO, G., DINTZIS, H. M., PARRISH, R. G., WYCKOFF, H. & PHILLIPS, D. C. 1958. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature,* 181**,** 662-6.

KIM, S., KIM, D., CHO, S. W., KIM, J. & KIM, J. S. 2014. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res,* 24**,** 1012-9.

KOONIN, E. V., MAKAROVA, K. S. & ZHANG, F. 2017. Diversity, classification and evolution of CRISPR-Cas systems. *Curr Opin Microbiol,* 37**,** 67-78.

KRISSINEL, E. & HENRICK, K. 2007. Inference of macromolecular assemblies from crystalline state. *J Mol Biol,* 372**,** 774-97.

LANCASTER, D., ELAM, S. & KAISER, A. B. 1989. Immunogenicity of the intradermal route of hepatitis B vaccination with the use of recombinant hepatitis B vaccine. *Am J Infect Control,* 17**,** 126-9.

LINTNER, N. G., KEROU, M., BRUMFIELD, S. K., GRAHAM, S., LIU, H., NAISMITH, J. H., SDANO, M., PENG, N., SHE, Q., COPIE, V., YOUNG, M. J., WHITE, M. F. & LAWRENCE, C. M. 2011. Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). *J Biol Chem,* 286**,** 21643-56.

MAKAROVA, K. S., ARAVIND, L., WOLF, Y. I. & KOONIN, E. V. 2011a. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol Direct,* 6**,** 38.

MAKAROVA, K. S., GRISHIN, N. V., SHABALINA, S. A., WOLF, Y. I. & KOONIN, E. V. 2006. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct,* 1**,** 7.

MAKAROVA, K. S., HAFT, D. H., BARRANGOU, R., BROUNS, S. J., CHARPENTIER, E., HORVATH, P., MOINEAU, S., MOJICA, F. J., WOLF, Y. I., YAKUNIN, A. F., VAN DER OOST, J. & KOONIN, E. V. 2011b. Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol,* 9**,** 467-77.

MAKAROVA, K. S. & KOONIN, E. V. 2015. Annotation and Classification of CRISPR-Cas Systems. *Methods Mol Biol,* 1311**,** 47-75.

MAKAROVA, K. S., WOLF, Y. I., ALKHNBASHI, O. S., COSTA, F., SHAH, S. A., SAUNDERS, S. J., BARRANGOU, R., BROUNS, S. J., CHARPENTIER, E., HAFT, D. H., HORVATH, P., MOINEAU, S., MOJICA, F. J., TERNS, R. M., TERNS, M. P., WHITE, M. F., YAKUNIN, A. F., GARRETT, R. A., VAN DER OOST, J., BACKOFEN, R. & KOONIN, E. V. 2015. An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol,* 13**,** 722-36.

MAKAROVA, K. S., WOLF, Y. I. & KOONIN, E. V. 2013. The basic building blocks and evolution of CRISPR-CAS systems. *Biochem Soc Trans,* 41**,** 1392-400.

MALI, P., YANG, L., ESVELT, K. M., AACH, J., GUELL, M., DICARLO, J. E., NORVILLE, J. E. & CHURCH, G. M. 2013. RNA-guided human genome engineering via Cas9. *Science,* 339**,** 823-6.

MARRAFFINI, L. A. & SONTHEIMER, E. J. 2008. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science,* 322**,** 1843-5.

MARRAFFINI, L. A. & SONTHEIMER, E. J. 2010a. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet,* 11**,** 181-90.

MARRAFFINI, L. A. & SONTHEIMER, E. J. 2010b. Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature,* 463**,** 568-71.

MCCOY, A. J., GROSSE-KUNSTLEVE, R. W., ADAMS, P. D., WINN, M. D., STORONI, L. C. & READ, R. J. 2007. Phaser crystallographic software. *J Appl Crystallogr,* 40**,** 658-674.

MCPHERSON, A. 2004. Introduction to protein crystallization. *Methods,* 34**,** 254-65.

MCREE, D., DAVID, P. 1999. *Practical Protein Crystallography,* San Diego, Elsevier.

MEDLOCK, A. E., CARTER, M., DAILEY, T. A., DAILEY, H. A. & LANZILOTTA, W. N. 2009. Product release rather than chelation determines metal specificity for ferrochelatase. *J Mol Biol,* 393**,** 308-19.

MOJICA, F. J., DIEZ-VILLASENOR, C., GARCIA-MARTINEZ, J. & SORIA, E. 2005. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol,* 60**,** 174-82.

MOJICA, F. J., DIEZ-VILLASENOR, C., SORIA, E. & JUEZ, G. 2000. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol Microbiol,* 36**,** 244-6.

MOJICA, F. J., FERRER, C., JUEZ, G. & RODRIGUEZ-VALERA, F. 1995. Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning. *Mol Microbiol,* 17**,** 85-93.

MOJICA, F. J., JUEZ, G. & RODRIGUEZ-VALERA, F. 1993. Transcription at different salinities of Haloferax mediterranei sequences adjacent to partially modified PstI sites. *Mol Microbiol,* 9**,** 613-21.

MULEPATI, S. & BAILEY, S. 2011. Structural and biochemical analysis of nuclease domain of clustered regularly interspaced short palindromic repeat (CRISPR)-associated protein 3 (Cas3). *J Biol Chem,* 286**,** 31896-903.

MULEPATI, S. & BAILEY, S. 2013. In vitro reconstitution of an Escherichia coli RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA target. *J Biol Chem,* 288**,** 22184-92.

MULEPATI, S., HEROUX, A. & BAILEY, S. 2014. Structural biology. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. *Science,* 345**,** 1479-84.

NAGEM, R. A., POLIKARPOV, I. & DAUTER, Z. 2003. Phasing on rapidly soaked ions. *Methods Enzymol,* 374**,** 120-37.

NISHIMASU, H., RAN, F. A., HSU, P. D., KONERMANN, S., SHEHATA, S. I., DOHMAE, N., ISHITANI, R., ZHANG, F. & NUREKI, O. 2014. Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell,* 156**,** 935-49.

NUMATA, T., INANAGA, H., SATO, C. & OSAWA, T. 2015. Crystal structure of the Csm3-Csm4 subcomplex in the type III-A CRISPR-Cas interference complex. *J Mol Biol,* 427**,** 259-73.

OSAWA, T., INANAGA, H., SATO, C. & NUMATA, T. 2015. Crystal Structure of the CRISPR-Cas RNA Silencing Cmr Complex Bound to a Target Analog. *Mol Cell,* 58**,** 418-30.

PENG, W., FENG, M., FENG, X., LIANG, Y. X. & SHE, Q. 2015. An archaeal CRISPR type III-B system exhibiting distinctive RNA targeting features and mediating dual RNA and DNA interference. *Nucleic Acids Res,* 43**,** 406-17.

PERUTZ, M. F., ROSSMANN, M. G., CULLIS, A. F., MUIRHEAD, H., WILL, G. & NORTH, A. C. 1960. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. *Nature,* 185**,** 416-22.

POLIKARPOV, I., PERLES, L. A., DE OLIVEIRA, R. T., OLIVA, G., CASTELLANO, E. E., GARRATT, R. C. & CRAIEVICH, A. 1998. Set-up and experimental parameters of the protein crystallography beamline at the Brazilian National Synchrotron Laboratory. *J Synchrotron Radiat,* 5**,** 72-6.

POLLOCK, D. P., KUTZKO, J. P., BIRCK-WILSON, E., WILLIAMS, J. L., ECHELARD, Y. & MEADE, H. M. 1999. Transgenic milk as a method for the production of recombinant antibodies. *J Immunol Methods,* 231**,** 147-57.

POURCEL, C., SALVIGNOL, G. & VERGNAUD, G. 2005. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology,* 151**,** 653-63.

PRASZKIER, J. & PITTARD, A. J. 2005. Control of replication in I-complex plasmids. *Plasmid,* 53**,** 97-112.

REEKS, J., GRAHAM, S., ANDERSON, L., LIU, H., WHITE, M. F. & NAISMITH, J. H. 2013a. Structure of the archaeal Cascade subunit Csa5: relating the small subunits of CRISPR effector complexes. *RNA Biol,* 10**,** 762-9.

REEKS, J., NAISMITH, J. H. & WHITE, M. F. 2013b. CRISPR interference: a structural perspective. *Biochem J,* 453**,** 155-66.

RICHTER, C., CHANG, J. T. & FINERAN, P. C. 2012. Function and regulation of clustered regularly interspaced short palindromic repeats (CRISPR) / CRISPR associated (Cas) systems. *Viruses,* 4**,** 2291-311.

ROBBINS, A. H., MCREE, D. E., WILLIAMSON, M., COLLETT, S. A., XUONG, N. H., FUREY, W. F., WANG, B. C. & STOUT, C. D. 1991. Refined crystal structure of Cd, Zn metallothionein at 2.0 A resolution. *J Mol Biol,* 221**,** 1269-93.

ROUILLON, C., ZHOU, M., ZHANG, J., POLITIS, A., BEILSTEN-EDMANDS, V., CANNONE, G., GRAHAM, S., ROBINSON, C. V., SPAGNOLO, L. & WHITE, M. F. 2013. Structure of the CRISPR interference complex CSM reveals key similarities with cascade. *Mol Cell,* 52**,** 124-34.

SAMAI, P., PYENSON, N., JIANG, W., GOLDBERG, G. W., HATOUM-ASLAN, A. & MARRAFFINI, L. A. 2015. Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. *Cell,* 161**,** 1164-74.

SANDER, J. D. & JOUNG, J. K. 2014. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat Biotechnol,* 32**,** 347-55.

SAPRANAUSKAS, R., GASIUNAS, G., FREMAUX, C., BARRANGOU, R., HORVATH, P. & SIKSNYS, V. 2011. The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic Acids Res,* 39**,** 9275-82.

SASHITAL, D. G., WIEDENHEFT, B. & DOUDNA, J. A. 2012. Mechanism of foreign DNA selection in a bacterial adaptive immune system. *Mol Cell,* 46**,** 606-15.

SEMENOVA, E., KUZNEDELOV, K., DATSENKO, K. A., BOUDRY, P. M., SAVITSKAYA, E. E., MEDVEDEVA, S., BELOGLAZOVA, N., LOGACHEVA, M., YAKUNIN, A. F. & SEVERINOV, K. 2015. The Cas6e ribonuclease is not required for interference and adaptation by the E. coli type I-E CRISPR-Cas system. *Nucleic Acids Res,* 43**,** 6049-61.

SHAH, S. A., ERDMANN, S., MOJICA, F. J. & GARRETT, R. A. 2013. Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biol,* 10**,** 891-9.

SHMAKOV, S., ABUDAYYEH, O. O., MAKAROVA, K. S., WOLF, Y. I., GOOTENBERG, J. S., SEMENOVA, E., MINAKHIN, L., JOUNG, J., KONERMANN, S., SEVERINOV, K., ZHANG, F. & KOONIN, E. V. 2015. Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. *Mol Cell,* 60**,** 385-97.

SIEVERS, F., WILM, A., DINEEN, D., GIBSON, T. J., KARPLUS, K., LI, W., LOPEZ, R., MCWILLIAM, H., REMMERT, M., SODING, J., THOMPSON, J. D. & HIGGINS, D. G. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol,* 7**,** 539.

SIEVERT, V., ERGIN, A. & BUSSOW, K. 2008. High throughput cloning with restriction enzymes. *Methods Mol Biol,* 426**,** 163-73.

SINKUNAS, T., GASIUNAS, G., FREMAUX, C., BARRANGOU, R., HORVATH, P. & SIKSNYS, V. 2011. Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J,* 30**,** 1335-42.

SORENSEN, H. P. & MORTENSEN, K. K. 2005. Advanced genetic strategies for recombinant protein expression in Escherichia coli. *J Biotechnol,* 115**,** 113-28.

SPILMAN, M., COCOZAKI, A., HALE, C., SHAO, Y., RAMIA, N., TERNS, R., TERNS, M., LI, H. & STAGG, S. 2013. Structure of an RNA silencing complex of the CRISPR-Cas immune system. *Mol Cell,* 52**,** 146-52.

STAALS, R. H., AGARI, Y., MAKI-YONEKURA, S., ZHU, Y., TAYLOR, D. W., VAN DUIJN, E., BARENDREGT, A., VLOT, M., KOEHORST, J. J., SAKAMOTO, K., MASUDA, A., DOHMAE, N., SCHAAP, P. J., DOUDNA, J. A., HECK, A. J., YONEKURA, K., VAN DER OOST, J. & SHINKAI, A. 2013. Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of Thermus thermophilus. *Mol Cell,* 52**,** 135-45.

STAALS, R. H., ZHU, Y., TAYLOR, D. W., KORNFELD, J. E., SHARMA, K., BARENDREGT, A., KOEHORST, J. J., VLOT, M., NEUPANE, N., VAROSSIEAU, K., SAKAMOTO, K., SUZUKI, T., DOHMAE, N., YOKOYAMA, S., SCHAAP, P. J., URLAUB, H., HECK, A. J., NOGALES, E., DOUDNA, J. A., SHINKAI, A. & VAN DER OOST, J. 2014. RNA targeting by the type III-A CRISPR-Cas Csm complex of Thermus thermophilus. *Mol Cell,* 56**,** 518-30.

SUN, P. D., RADAEV, S. & KATTAH, M. 2002. Generating isomorphous heavy-atom derivatives by a quick-soak method. Part I: test cases. *Acta Crystallogr D Biol Crystallogr,* 58**,** 1092-8.

TAMULAITIS, G., KAZLAUSKIENE, M., MANAKOVA, E., VENCLOVAS, C., NWOKEOJI, A. O., DICKMAN, M. J., HORVATH, P. & SIKSNYS, V. 2014. Programmable RNA shredding by the type III-A CRISPR-Cas system of Streptococcus thermophilus. *Mol Cell,* 56**,** 506-17.

TAMULAITIS, G., VENCLOVAS, C. & SIKSNYS, V. 2017. Type III CRISPR-Cas Immunity: Major Differences Brushed Aside. *Trends Microbiol,* 25**,** 49-61.

TAYLOR, D. W., ZHU, Y., STAALS, R. H., KORNFELD, J. E., SHINKAI, A., VAN DER OOST, J., NOGALES, E. & DOUDNA, J. A. 2015. Structural biology. Structures of the CRISPR-Cmr complex reveal mode of RNA target positioning. *Science,* 348**,** 581-5.

TAYLOR, G. L. 2010. Introduction to phasing. *Acta Crystallogr D Biol Crystallogr,* 66**,** 325-38.

TRITOS, N. A. & MANTZOROS, C. S. 1998. Recombinant human growth hormone: old and novel uses. *Am J Med,* 105**,** 44-57.

TROWITZSCH, S., BIENIOSSEK, C., NIE, Y., GARZONI, F. & BERGER, I. 2010. New baculovirus expression tools for recombinant protein complex production. *J Struct Biol,* 172**,** 45-54.

VAN DER OOST, J., WESTRA, E. R., JACKSON, R. N. & WIEDENHEFT, B. 2014. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nat Rev Microbiol,* 12**,** 479-92.

VESTERGAARD, G., GARRETT, R. A. & SHAH, S. A. 2014. CRISPR adaptive immune systems of Archaea. *RNA Biol,* 11**,** 156-67.

WANG, R. & LI, H. 2012. The mysterious RAMP proteins and their roles in small RNA-based immunity. *Protein Sci,* 21**,** 463-70.

WEI, Y., TERNS, R. M. & TERNS, M. P. 2015. Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes Dev,* 29**,** 356-61.

WIEDENHEFT, B., LANDER, G. C., ZHOU, K., JORE, M. M., BROUNS, S. J., VAN DER OOST, J., DOUDNA, J. A. & NOGALES, E. 2011a. Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature,* 477**,** 486-9.

WIEDENHEFT, B., VAN DUIJN, E., BULTEMA, J. B., WAGHMARE, S. P., ZHOU, K., BARENDREGT, A., WESTPHAL, W., HECK, A. J., BOEKEMA, E. J., DICKMAN, M. J. & DOUDNA, J. A. 2011b. RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc Natl Acad Sci U S A,* 108**,** 10092-7.

WILLIAMS, S. G., CRANENBURGH, R. M., WEISS, A. M., WRIGHTON, C. J., SHERRATT, D. J. & HANAK, J. A. 1998. Repressor titration: a novel system for selection and stable maintenance of recombinant plasmids. *Nucleic Acids Res,* 26**,** 2120-4.

YOGAVEL, M., NITHYA, N., SUZUKI, A., SUGIYAMA, Y., YAMANE, T., VELMURUGAN, D. & SHARMA, A. 2010. Structural analysis of actinidin and a comparison of cadmium and sulfur anomalous signals from actinidin crystals measured using in-house copper- and chromium-anode X-ray sources. *Acta Crystallogr D Biol Crystallogr,* 66**,** 1323-33.

ZETSCHE, B., GOOTENBERG, J. S., ABUDAYYEH, O. O., SLAYMAKER, I. M., MAKAROVA, K. S., ESSLETZBICHLER, P., VOLZ, S. E., JOUNG, J., VAN DER OOST, J., REGEV, A., KOONIN, E. V. & ZHANG, F. 2015. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell,* 163**,** 759-71.

ZHAO, H., SHENG, G., WANG, J., WANG, M., BUNKOCZI, G., GONG, W., WEI, Z. & WANG, Y. 2014. Crystal structure of the RNA-guided immune surveillance Cascade complex in Escherichia coli. *Nature,* 515**,** 147-50.

ZHENG, X., YANG, S., ZHANG, D., ZHONG, Z., TANG, X., DENG, K., ZHOU, J., QI, Y. & ZHANG, Y. 2016. Effective screen of CRISPR/Cas9-induced mutants in rice by single-strand conformation polymorphism. *Plant Cell Rep,* 35**,** 1545-54.

8 ANEXOS

STRUCTURAL BIOLOGY COMMUNICATIONS

ISSN 2053-230X

Purification, crystallization, crystallographic analysis and phasing of the CRISPR-associated protein Csm2 from *Thermotoga maritima*

Gloria Gallo,[a] Gilles Augusto,[a] Giulliana Rangel,[a] André Zelanis,[a] Marcelo A. Mori,[b] Cláudia Barbosa Campos[a] and Martin Würtele[a]*

[a]Department of Science and Technology, Federal University of São Paulo, Rua Talim 330, 12231-280 São José dos Campos-SP, Brazil, and [b]Department of Biophysics, Federal University of São Paulo, Rua Botucatu 862, 04023-062 São Paulo-SP, Brazil. *Correspondence e-mail: martin.wurtele@unifesp.br

Received 22 June 2015
Accepted 6 August 2015

Edited by M. S. Weiss, Helmholtz-Zentrum Berlin für Materialien und Energie, Germany

Keywords: RNA-guided interference; CRISPR–Cas; Csm2; *Thermotoga maritima*.

The clusters of regularly interspaced short palindromic repeats (CRISPR)–CRISPR-associated proteins (Cas) system consists of an intriguing machinery of proteins that confer bacteria and archaea with immunity against phages and plasmids *via* an RNA-guided interference mechanism. Here, the cloning, recombinant expression in *Escherichia coli* BL21 (DE3), purification, crystallization and preliminary X-ray diffraction analysis of Csm2 from *Thermotoga maritima* are reported. Csm2 is thought to be a component of an important protein complex of the type IIIA CRISPR–Cas system, which is involved in the CRISPR–Cas RNA-guided interference pathway. The structure of Csm2 was solved *via* cadmium single-wavelength anomalous diffraction (Cd-SAD) phasing. Owing to its involvement in the CRISPR–Cas system, the crystal structure of this protein could be of importance in elucidating the mechanism of type IIIA CRISPR–Cas systems in bacteria and archaea.

1. Introduction

Recent reports have characterized the bacterial clusters of regularly interspaced short palindromic repeats (CRISPR)–CRISPR-associated proteins (Cas) system as a prokaryotic adaptive immune system based on RNA-guided interference (Mojica *et al.*, 2005; Bolotin *et al.*, 2005; Pourcel *et al.*, 2005; Makarova *et al.*, 2006; Barrangou *et al.*, 2007). The CRISPR–Cas system is effective in protecting cells against nucleotide uptake from bacteriophages and plasmids (Barrangou *et al.*, 2007; Marraffini & Sontheimer, 2008). In summary, foreign DNA is integrated into the CRISPR elements of the prokaryotic host genome (Pourcel *et al.*, 2005; Bolotin *et al.*, 2005; Barrangou *et al.*, 2007) and transcripts of this DNA in the form of processed crRNA (CRISPR RNA) are involved in an RNA-guided interference mechanism to degrade target DNA or RNA (Brouns *et al.*, 2008; Wiedenheft *et al.*, 2009; Hale *et al.*, 2009; Marraffini & Sontheimer, 2010).

Depending on the occurrence of three signature proteins, Cas3, Cas9 and Cas10, CRISPR–Cas systems are classified as type I, type II and type III systems, respectively (Makarova, Aravind *et al.*, 2011; Makarova, Haft *et al.*, 2011). Type III CRISPR–Cas systems are further subdivided into two different subtypes called IIIA and IIIB. Whereas type I and type II CRISPR–Cas systems target DNA (Garneau *et al.*, 2010; Sinkunas *et al.*, 2011; Gasiunas *et al.*, 2012), type III systems are thought to target both DNA and RNA (Marraffini & Sontheimer, 2008; Hale *et al.*, 2009; Staals *et al.*, 2013, 2014;

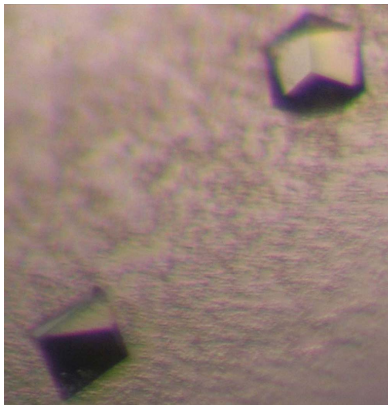


Table 1
Macromolecule-production information.

Source organism	*T. maritima*
DNA source	*T. maritima*
Forward primer (BamHI restriction site)	5′-GGCCGGGGATCCGCAGTTTCTCAGGGTGTTTC
Reverse primer (HindIII restriction site)	5′-GGCCGGAAGCTTATCTTCTGGCTGTCCTTCT-GTTG
Cloning vector	pQtev
Expression vector	pQtev
Expression host	*E. coli* BL21 (DE3)
Complete amino-acid sequence of the construct produced	GSAVSQGVSLKEDLKDLVRKAEEIGRELSGKLKT-NQLRKFHGHLTKIWSNYIYKKKDYRDNPEKFN-EEILNELHFMKIFLAYQVGRDIEGISELKEIL-EPLIDEIKTPDEFEKFKKFYDAILAYHKFHSE-SEKSNRRTARR

Taylor *et al.*, 2015; Samai *et al.*, 2015). The type IIIA system includes Csm Cas proteins (Makarova, Haft *et al.*, 2011), which were identified as being expressed in *Mycobacterium tuberculosis* (Haft *et al.*, 2005). The type IIIB system includes Cmr proteins (Hale *et al.*, 2009). Processing and target recognition by crRNA is carried out by several Cas proteins, which usually form a complex of ribonucleoproteins (RNPs) like the so-called Cascade complex in type I systems (Brouns *et al.*, 2008) and the Cmr CRISPR RNP (crRNP) complex in type IIIB systems (Hale *et al.*, 2012; Staals *et al.*, 2013). Electron-microscopy data and protein crystallographic studies have revealed the composition and function of the Cascade (Jore *et al.*, 2011; Wiedenheft *et al.*, 2011; Jackson *et al.*, 2014) and the Cmr crRNP complexes (Staals *et al.*, 2013; Osawa *et al.*, 2015) in great detail. Furthermore, electron-microscopy and mass-spectrometric data have defined the Csm crRNP complex from *Sulfolobus solfataricus*, which is formed by up to seven different proteins (Rouillon *et al.*, 2013). Similar experiments have defined a homologous complex in *Thermus thermophilus* that consists of five different proteins: Csm1–Csm5 (Staals *et al.*, 2014). Cmr, Csm and Cascade complexes show overall structural similarity, possibly owing to the fact that they are composed of structurally related subunits (Rouillon *et al.*, 2013). To obtain further insight into the Csm crRNP complex, we have cloned, expressed, purified, crystallized and solved the structure of the Csm2 protein from *Thermotoga maritima* MSB8.

2. Materials and methods

2.1. Cloning and expression

Full-length Csm2 (GenBank entry AKE29563.1) was amplified *via* PCR with primer overhangs from total genomic extracts of *T. maritima* MSB8 (Huber *et al.*, 1986) obtained from DSMZ (German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany). DNA oligonucleotides (Exxtend, Paulínia, Brazil) are described in Table 1. PCR fragments were cloned into the pQtev His-tag expression vector (Protein Structure Factory, Berlin, Germany) using BamHI and HindIII cloning sites. The construct was verified by sequencing (Proteobras, Paulínia, Brazil). Csm2 was expressed in LB medium for 6 h at 37°C in *Escherichia coli* strain BL21 (DE3) after induction with 1 m*M* β-D-1-thio-galactopyranoside (IPTG).

Table 2
Crystallization.

Method	Hanging drop
Plate type	24-well VDX plate
Temperature (K)	293.15
Protein concentration (mg ml^{-1})	15
Buffer composition of protein solution	25 m*M* Tris–HCl pH 7.5, 100 m*M* NaCl, 3 m*M* DTT, 10% glycerol
Composition of reservoir solution	100 m*M* sodium acetate pH 4.6, 100 m*M* $CdCl_2$, 21% PEG 400
Volume and ratio of drop	4 µl, 1:1 ratio
Volume of reservoir (ml)	1

2.2. Protein purification

The cells were harvested by centrifugation and resuspended in lysis buffer consisting of 50 m*M* Tris–HCl pH 7.5, 400 m*M* NaCl, 10 m*M* imidazole, 1 m*M* PMSF, 5 m*M* β-mercaptoethanol. The cells were lysed using an M-110L high-pressure homogenizer (Microfluidics, Westwood, USA). After separation of the soluble fraction by centrifugation, the protein was purified by affinity chromatography using a 5 ml HisTrap Sepharose column (GE Healthcare) on an ÄKTAprime Plus liquid-chromatography system (GE Healthcare). After elution with imidazole and dialysis in 50 m*M* Tris–HCl pH 7.5, 400 m*M* NaCl, 3 m*M* DTT, 10% glycerol, the His tag was cleaved using TEV (*Tobacco etch virus*) protease (Blommel & Fox, 2007). After a subsequent affinity-chromatography step to remove the protease and the His tag, Csm2 was concentrated using Amicon Ultra-15 centrifugal filters (Millipore) and submitted to gel-filtration chromatography on a HiLoad 26/600 Superdex 75 prep-grade column with a nominal volume of 320 ml (GE Healthcare) in buffer consisting of 25 m*M* Tris–HCl pH 7.5, 100 m*M* NaCl, 3 m*M* DTT, 10% glycerol. Finally, the protein was concentrated to 15 mg ml^{-1}. The molecular mass of native Csm2 (650 ng) was confirmed by electrospray-ionization (ESI) mass spectrometry using a hybrid quadrupole (Q)-IM-ToF MS instrument (Synapt G2 HDMS mass spectrometer, Waters). Deconvolution of the MS protein spectrum yielded an isotope-averaged molecular mass of 16 732.78 Da, which is similar to the value obtained by SDS–PAGE using SeeBlue Plus2 Pre-stained protein standard as a marker (Invitrogen, Life Technologies).

2.3. Crystallization

Csm2 was crystallized by the hanging-drop method in 24-well plates after screening trials using Crystal Screen 2 (Hampton Research, Aliso Viejo, USA) at 20°C. Initial screening conditions were optimized, leading to optimal crystallization conditions consisting of 100 m*M* cadmium chloride, 21% PEG 400, 100 m*M* sodium acetate buffer pH 4.6, as described in Table 2.

2.4. Data collection and processing

Diffraction data were collected on the MX-1 beamline at the Brazilian Synchrotron Light Laboratory (LNLS, Campinas,

Table 3
Data collection and processing.

Values in parentheses are for the outer shell.

Diffraction source	MX-1, LNLS
Wavelength (Å)	1.458
Temperature (K)	100
Detector	MAR CCD 165
Crystal-to-detector distance (mm)	120
Rotation range per image (°)	1
Data range (images)	1–190
Space group	$P3_121$
a, b, c (Å)	77, 77, 160
α, β, γ (°)	90, 90, 120
Mosaicity (°)	0.2
Resolution range (Å)	50–2.9 (3.08–2.90)
Total No. of reflections	143907 (22874)
No. of unique reflections	23467 (3813)
Completeness (%)	99.9 (99.6)
Multiplicity	6.13 (5.99)
$\langle I/\sigma(I)\rangle$	13.77 (2.62)
R_{meas} (%)	11.0 (81.3)
Overall B factor from Wilson plot (Å^2)	81.2

Brazil; Polikarpov *et al.*, 1998) after flash-cooling in liquid nitrogen using crystallization buffer supplemented with 12% glycerol. A highly redundant data set was collected at a wavelength of 1.458 Å. Crystallographic data were processed with *XDS* (Kabsch, 2010) and *Adxv* (Arvai, 2015). The structure of Csm2 was solved by single-wavelength anomalous diffraction (SAD) using *Phaser* (McCoy *et al.*, 2007) within the *AutoSol* module of *PHENIX* (Adams *et al.*, 2010), using the fact that the crystallization conditions contained cadmium ions. Electron-density maps were inspected using *Coot* (Emsley *et al.*, 2010) and figures were rendered using *PyMOL* (DeLano, 2002).

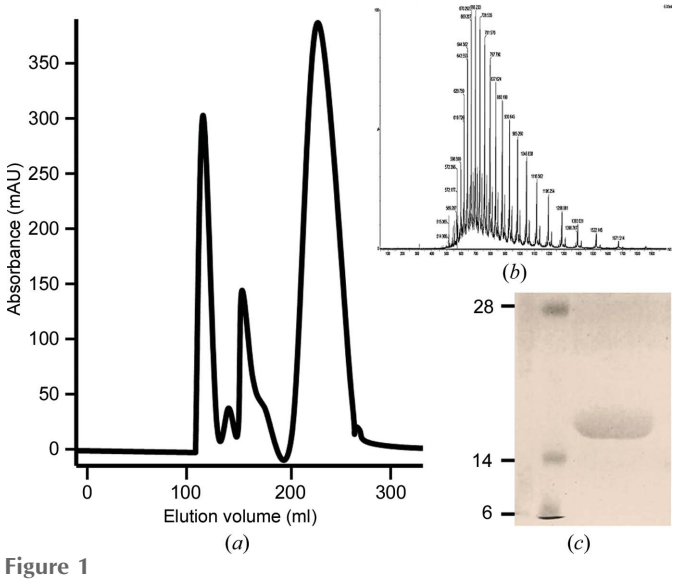
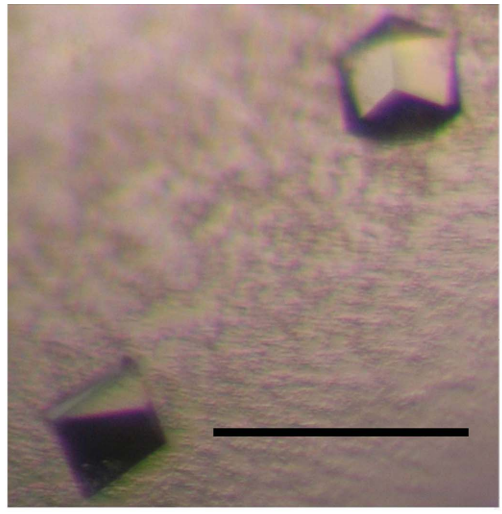


Figure 1
Purification of Csm2. (*a*) Gel-filtration chromatography profile of Csm2 using a HiLoad 26/600 Superdex 75 prep-grade column (GE Healthcare). The last peak corresponds to the crystallized Csm2. (*b*) Molecular-mass determination of native Csm2 by mass spectrometry. Deconvolution of the MS protein spectrum yielded an isotope-averaged molecular mass of 16 732.78 Da. (*c*) SDS–PAGE of Csm2 after gel filtration using SeeBlue Plus2 Pre-stained protein standard (Invitrogen, Life Technologies) as a marker (labelled in kDa).

3. Results and discussion

In recent years, several outstanding crystal structures of RNA-processing enzymes and protein complexes have been obtained using isoforms from thermophilic bacteria (Wang *et al.*, 2008; Mulepati & Bailey, 2011; Lintner *et al.*, 2011; Cocozaki *et al.*, 2012; Osawa *et al.*, 2013). To obtain further data for the Csm CRISPR–Cas type IIIA crRNP complex, we have cloned, expressed, purified and crystalized a major protein from this complex, Csm2.

T. maritima MS8 Csm2 was expressed in soluble form in *E. coli* strain BL21 (DE3) in relatively high quantities of up to 30 mg per litre of culture. However, the protein could initially not be concentrated owing to precipitation. This problem was



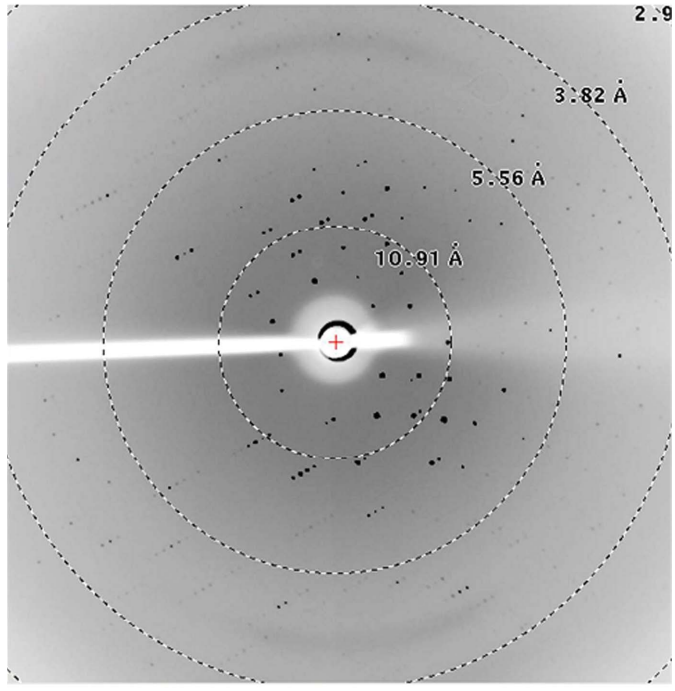


Figure 2
Csm2 crystals and X-ray diffraction pattern. (*a*) Representative native crystals of Csm2; scale bar = 0.5 mm. (*b*) Diffraction image of a Csm2 crystal.

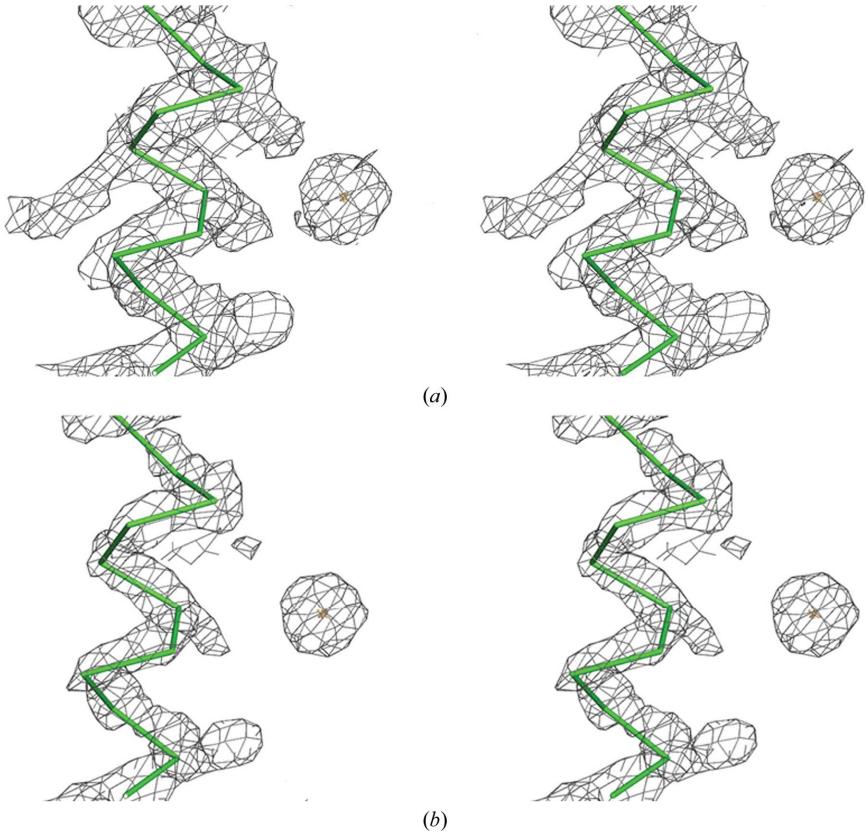


Figure 3
Electron-density maps. Stereoview of the *AutoSol* density-modified electron-density map of the Csm2 crystal structure contoured at a 1.5σ (*a*) and 3σ (*b*) level showing a cadmium ion (yellow) linked to an α-helix with a C^{α} trace (green).

circumvented by using glycerol (10%) and high concentrations of NaCl (up to 400 m*M*) during all purification steps. To improve the chance of crystallization, after purifying the protein the concentration of NaCl in the final gel-filtration step was reduced to 100 m*M* (Fig. 1). The protein could then be concentrated to 15 mg ml^{-1}. Small monocrystals obtained in initial screens could be improved in size through refinement of the crystallization conditions, as shown in Fig. 2.

A highly redundant data set was collected from one of these crystals on the MX-1 beamline at the Brazilian Synchrotron Light Laboratory (LNLS), Campinas, Brazil. Crystallographic data for this crystal are given in Table 3. The crystals tested diffracted to a limited resolution of 2.9 Å in space group $P3_121$ and contained an estimated three subunits in the asymmetric unit (Matthews coefficient of 2.7 Å^3 Da^{-1} and 54% solvent content). As there are no known close structural homologues of Csm2 in the PDB, the structure of Csm2 could not be solved by molecular replacement. However, as our crystals grew at high concentrations (100 m*M*) of $CdCl_2$, we hypothesized that Cd^{2+} ions possibly formed strong linkages with the numerous charged amino acids in Csm2. As several structures have been solved using Cd-SAD and/or S-SAD (Yogavel *et al.*, 2010; Robbins *et al.*, 1991; Medlock *et al.*, 2009), Cd-SAD was used together with the *PHENIX* software package (Adams *et al.*, 2010) to try to solve the structure. Indeed, the measured data set indicated the presence of an anomalous signal, quantified by an overall anomalous correlation of 31% and an overall mean anomalous difference in units of standard deviation (SigAno in *XDS*) of 1.085. Consequently, *phenix.hyss* was able to detect six cadmium ions. Automatic refinement and phasing with *Phaser* and subsequent phase density modifications then led to the first experimental electron-density maps. This solution was characterized by an overall figure of merit of 0.319, a correlation of local r.m.s. density of 0.82 and a meaningful map skew of 0.10.

Most importantly, the experimental electron-density maps after density modification promptly showed meaningful electron density, and several α-helical structures could be detected in these maps (Fig. 3). However, the resolution of the data set has thus far hindered the assignment of amino-acid side chains. Improvement of the diffraction resolution will thus be necessary in order to build and refine an unambiguous model.

As Csm2 is a homologue of *S. solfataricus* SSo1424 (18% sequence identity) and *T. thermophilus* Csm2 (22% sequence identity), which play important roles in the assembly of Csm crRNP complexes (Rouillon *et al.*, 2013; Staals *et al.*, 2014), the results described here have an impact in helping to understand the structure and function of Csm crRNP complexes of type IIIA CRISPR–Cas systems.

Acknowledgements

This work was supported by research grants 11/50963-4 from FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo, São Paulo Research Foundation), 480411/2011-5 from

CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, National Council for Scientific and Technological Development, Brazil) and 448833/2014-0 from CNPq. The authors thank the LNLS beamline staff for help with the measurements and Dr A. Tashima from the proteomics laboratory of UNIFESP for help with the MS measurements.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.

Arvai, A. (2015). *Adxv – A Program to Display X-ray Diffraction Images.* http://www.scripps.edu/tainer/arvai/adxv.html.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A. & Horvath, P. (2007). *Science*, **315**, 1709–1712.

Blommel, P. G. & Fox, B. G. (2007). *Protein Expr. Purif.* **55**, 53–68.

Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. (2005). *Microbiology*, **151**, 2551–2561.

Brouns, S. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J., Snijders, A. P., Dickman, M. J., Makarova, K. S., Koonin, E. V. & van der Oost, J. (2008). *Science*, **321**, 960–964.

Cocozaki, A. I., Ramia, N. F., Shao, Y., Hale, C. R., Terns, R. M., Terns, M. P. & Li, H. (2012). *Structure*, **20**, 545–553.

DeLano, W. L. (2002). *PyMOL*. http://www.pymol.org/.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.

Garneau, J. E., Dupuis, M.-È., Villion, M., Romero, D. A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A. H. & Moineau, S. (2010). *Nature (London)*, **468**, 67–71.

Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. (2012). *Proc. Natl Acad. Sci. USA*, **109**, E2579–E2586.

Haft, D. H., Selengut, J., Mongodin, E. F. & Nelson, K. E. (2005). *PLoS Comput. Biol.* **1**, e60.

Hale, C. R., Majumdar, S., Elmore, J., Pfister, N., Compton, M., Olson, S., Resch, A. M., Glover, C. V., Graveley, B. R., Terns, R. M. & Terns, M. P. (2012). *Mol. Cell*, **45**, 292–302.

Hale, C. R., Zhao, P., Olson, S., Duff, M. O., Graveley, B. R., Wells, L., Terns, R. M. & Terns, M. P. (2009). *Cell*, **139**, 945–956.

Huber, R. L., Langworthy, T. A., König, H., Thomm, M., Woese, C. R., Sleytr, U. B. & Stetter, K. O. (1986). *Arch. Microbiol.* **144**, 324–333.

Jackson, R. N., Golden, S. M., van Erp, P. B., Carter, J., Westra, E. R., Brouns, S. J., van der Oost, J., Terwilliger, T. C., Read, R. J. & Wiedenheft, B. (2014). *Science*, **345**, 1473–1479.

Jore, M. M. *et al.* (2011). *Nature Struct. Mol. Biol.* **18**, 529–536.

Kabsch, W. (2010). *Acta Cryst.* D**66**, 125–132.

Lintner, N. G., Kerou, M., Brumfield, S. K., Graham, S., Liu, H., Naismith, J. H., Sdano, M., Peng, N., She, Q., Copié, V., Young, M. J., White, M. F. & Lawrence, C. M. (2011). *J. Biol. Chem.* **286**, 21643–21656.

Makarova, K. S., Aravind, L., Wolf, Y. I. & Koonin, E. V. (2011). *Biol. Direct*, **6**, 38.

Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I. & Koonin, E. V. (2006). *Biol. Direct*, **1**, 7.

Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F. J., Wolf, Y. I., Yakunin, A. F., van der Oost, J. & Koonin, E. V. (2011). *Nature Rev. Microbiol.* **9**, 467–477.

Marraffini, L. A. & Sontheimer, E. J. (2008). *Science*, **322**, 1843–1845.

Marraffini, L. A. & Sontheimer, E. J. (2010). *Nature Rev. Genet.* **11**, 181–190.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.

Medlock, A. E., Carter, M., Dailey, T. A., Dailey, H. A. & Lanzilotta, W. N. (2009). *J. Mol. Biol.* **393**, 308–319.

Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. (2005). *J. Mol. Evol.* **60**, 174–182.

Mulepati, S. & Bailey, S. (2011). *J. Biol. Chem.* **286**, 31896–31903.

Osawa, T., Inanaga, H. & Numata, T. (2013). *J. Mol. Biol.* **425**, 3811–3823.

Osawa, T., Inanaga, H., Sato, C. & Numata, T. (2015). *Mol. Cell*, **58**, 418–430.

Polikarpov, I., Perles, L. A., de Oliveira, R. T., Oliva, G., Castellano, E. E., Garratt, R. C. & Craievich, A. (1998). *J. Synchrotron Rad.* **5**, 72–76.

Pourcel, C., Salvignol, G. & Vergnaud, G. (2005). *Microbiology*, **151**, 653–663.

Robbins, A. H., McRee, D. E., Williamson, M., Collett, S. A., Xuong, N.-H., Furey, W. F., Wang, B.-C. & Stout, C. D. (1991). *J. Mol. Biol.* **221**, 1269–1293.

Rouillon, C., Zhou, M., Zhang, J., Politis, A., Beilsten-Edmands, V., Cannone, G., Graham, S., Robinson, C. V., Spagnolo, L. & White, M. F. (2013). *Mol. Cell*, **52**, 124–134.

Samai, P., Pyenson, N., Jiang, W., Goldberg, G. W., Hatoum-Aslan, A. & Marraffini, L. A. (2015). *Cell*, **161**, 1164–1174.

Sinkunas, T., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P. & Siksnys, V. (2011). *EMBO J.* **30**, 1335–1342.

Staals, R. H. *et al.* (2013). *Mol. Cell*, **52**, 135–145.

Staals, R. H. *et al.* (2014). *Mol. Cell*, **56**, 518–530.

Taylor, D. W., Zhu, Y., Staals, R. H., Kornfeld, J. E., Shinkai, A., van der Oost, J., Nogales, E. & Doudna, J. A. (2015). *Science*, **348**, 581–585.

Wang, Y., Juranek, S., Li, H., Sheng, G., Tuschl, T. & Patel, D. J. (2008). *Nature (London)*, **456**, 921–926.

Wiedenheft, B., Lander, G. C., Zhou, K., Jore, M. M., Brouns, S. J., van der Oost, J., Doudna, J. A. & Nogales, E. (2011). *Nature (London)*, **477**, 486–489.

Wiedenheft, B., Zhou, K., Jinek, M., Coyle, S. M., Ma, W. & Doudna, J. A. (2009). *Structure*, **17**, 904–912.

Yogavel, M., Nithya, N., Suzuki, A., Sugiyama, Y., Yamane, T., Velmurugan, D. & Sharma, A. (2010). *Acta Cryst.* D**66**, 1323–1333.

The FEBS Journal

FEBS PRESS science publishing by scientists

Structural basis for dimer formation of the CRISPR-associated protein Csm2 of *Thermotoga maritima*

Gloria Gallo[1], Gilles Augusto[1], Giulliana Rangel[1], André Zelanis[1,2], Marcelo A. Mori[3], Cláudia B. Campos[1] and Martin Würtele[1]

1 Department of Science and Technology, Federal University of São Paulo, São José dos Campos, Brazil
2 Applied Toxinology Laboratory – LETA and Center of Toxins, Immune-response and Cell Signaling – CeTICS, Instituto Butantan, São Paulo, Brazil
3 Department of Biophysics, Federal University of São Paulo, São Paulo, Brazil

Keywords
Cd-SAD; CRISPR-Cas; crRNP; Csm2; *Thermotoga maritima*

Correspondence
M. Würtele, Department of Science and Technology, Federal University of São Paulo, Rua Talim 330, São José dos Campos 12231-280, Brazil
Fax: +55 12 39218857
Tel: +55 12 33099698
E-mail: martin.wurtele@unifesp.br

(Received 30 September 2015, revised 30 November 2015, accepted 3 December 2015)

doi:10.1111/febs.13621

The clusters of regularly interspaced short palindromic repeats (CRISPR) and the Cas (CRISPR-associated) proteins form an adaptive immune system in bacteria and archaea that evolved as an RNA-guided interference mechanism to target and degrade foreign genetic elements. In the so-called type IIIA CRISPR-Cas systems, Cas proteins from the Csm family form a complex of RNPs that are involved in surveillance and targeting tasks. In the present study, we report the crystal structure of *Thermotoga maritima* Csm2. This protein is considered to assemble into the helically shaped Csm RNP complex in a site opposite to the CRISPR RNA binding backbone. Csm2 was solved via cadmium single wavelength anomalous diffraction phasing at 2.4 Å resolution. The structure reveals that Csm2 is composed of a large 42 amino-acid long α-helix flanked by three shorter α-helices. The structure also shows that the protein is capable of forming dimers mainly via an extensive contact surface conferred by its long α-helix. This interaction is further stabilized by the N-terminal helix, which is inserted into the C-terminal helical portion of the adjacent subunit. The dimerization of Csm2 was additionally confirmed by size exclusion chromatography of the pure recombinant protein followed by MS analysis of the eluted fractions. Because of its role in the assembly and functioning of the Csm CRISPR RNP complex, the crystal structure of Csm2 is of great importance for clarifying the mechanism of action of the subtype IIIA CRISPR-Cas system, as well as the similarities and diversities between the different CRISPR-Cas system.

Database
The structure of *Thermotoga maritima* Csm2 has been deposited in the Protein Data Bank under accession code 5AN6.

Introduction

The bacterial and archaeal clusters of regularly interspaced short palindromic repeats (CRISPR) consist of an array of short genomic repeat sequences separated by spacers and located adjacently to a cluster of genes expressing the so-called CRISPR-associated (Cas) proteins. Remarkably, the CRISPR-Cas system has been

Abbreviations
Cas, CRISPR associated; Cd-SAD, cadmium single wavelength anomalous diffraction; CRISPR, clusters of regularly interspaced short palindromic repeats; crRNA, CRISPR RNA; crRNP, CRISPR ribonucleoproteins; LNLS, Brazilian Synchrotron Light Laboratory; PDB, Protein Data Bank.

recently characterized as an RNA-guided interference based form of a rudimental adaptive immune system of prokaryotes [1–5] that acts to protect cells against the incorporation of bacteriophages or plasmids [5,6]. CRISPR-Cas is considered to function by integrating foreign DNA into the CRISPR spacer sequences [2,3,5], which are transcribed and processed to form CRISPR RNA (crRNA). In turn, crRNA is involved in an RNA-guided interference mechanism that targets nonhost DNA or RNA [7–10].

Highlighting their importance for the survival of their hosts, CRISPR-Cas systems are found in approximately half of all bacterial and in more than 80% of all known archaeal genomes [11]. CRISPR-Cas are classified as type I, type II and type III systems depending on the occurrence of the three signature proteins, Cas3, Cas9 and Cas10, respectively [12,13]. The type III CRISPR-Cas systems are further subdivided into two different subtypes, called III-A and III-B. Type I and type II CRISPR-Cas systems are known to target DNA [14–16]. Type III systems are assumed to target both DNA and RNA [6,8,17–20]. The type III-A system includes Csm Cas-proteins [13], whereas the type III-B system includes Cmr proteins [8].

In CRISPR-Cas systems, target recognition is carried out by Cas proteins, which form a complex of RNPs, similar to the so-called Cascade complex in type I systems [7], the Csm crRNP (CRISPR RNP) complex in type III-A systems [18] and the Cmr crRNP complex in type III-B systems [17,21]. Electron microscopy data and protein crystallographic studies have revealed in detail the composition and function of the Cascade [22–26] and the Cmr crRNP complexes [17,27]. Furthermore, electron microscopy and MS data have defined the Csm crRNP complex from *Sulfolobus solfataricus*, which is formed by up to seven different proteins [28] and a homologous complex in *Thermus thermophilus* that consists of five different proteins: Csm1–5 [18]. Cmr, Csm and Cascade complexes show overall structural similarity because they are formed by structurally and functionally related subunits [27,28]. In the present study, aiming to obtain more more information about the Csm crRNP complex, we cloned, expressed, purified, crystallized and solved the structure of the Csm2 protein of *Thermotoga maritima* MSB8.

Results and Discussion

Crystallization and refinement

Recent progress in understanding the structural mechanism of crRNP target oligonucleotide degrading complexes has advanced considerably in recent years [17,18,20,24–30]. To obtain more data from the Csm CRISPR-Cas type III-A crRNP complex, we have cloned, expressed, purified, crystallized and solved the structure of Csm2, which is a major protein of this complex.

T. maritima MSB8 Csm2 was expressed recombinantly in *Escherichia coli* strain BL21(DE3) and crystallized in space group $P3_121$. An initial data set of limited resolution (2.9 Å) of these crystals was originally measured at the MX1 beamline of the Brazilian Synchrotron Light Laboratory (LNLS, Campinas, Brazil) [31]. Because the crystal conditions contained 100 mM $CdCl_2$, we were recently [31] able to solve this structure using cadmium single wavelength anomalous diffraction (Cd-SAD) with the Autosol module of the software package PHENIX [32]. To be able to assign most of the amino acid chains in the electron density maps, additional crystals were measured, allowing us to obtain a single crystallographic dataset of a crystal with 2.4 Å resolution. Data collection statistics of this crystal are summarized in Table 1.

At the resolution of 2.4 Å, a final model of the Csm2 structure could be fitted and refined. The final model, which is characterized by a crystallographic R_{work} factor of 0.2098 and an R_{free} factor of 0.2488, contains three subunits of *T. maritima* Csm2 in the

Table 1. Crystallographic data collection statistics. Values for the outer shell are given in parentheses.

	Solution dataset [31]	Refinement dataset
Diffraction source	MX-1, LNLS	MX-2, LNLS
Wavelength (Å)	1.458	1.458
Temperature (K)	100	100
Detector	MARCCD165	PILATUS2M
Crystal-detector distance (mm)	120	170
Rotation range per image (°)	1	1
Data range	1–190	1–126
Space group	$P3_121$	$P3_121$
a, *b*, *c* (Å)	77, 77, 160	77, 77, 160
α, β, γ (°)	90,90, 120	90,90, 120
Mosaicity (°)	0.2	0.2
Resolution range (Å)	50–2.9 (3.08–2.9)	50–2.4 (2.5–2.4)
Total number of reflections	143 907 (22 874)	135 368 (16 640)
Number of unique reflections	23 467 (3813)	21 688 (3305)
Completeness (%)	99.9 (99.6)	98.3 (94.9)
Redundancy	6.14 (5.99)	6.24 (5.03)
$\langle I/\sigma(I)\rangle$	13.77 (2.62)	12.63 (3.64)
R_{meas} (%)	11.0 (81.31)	12.9 (83.9)

Table 2. Crystallographic structure refinement statistics. Values for the outer shell are given in parentheses.

Resolution range (Å)	41.65–2.40 (2.51–2.40)
Completeness (%)	98.16 (93)
$\langle I/\sigma(I)\rangle$	12.63 (3.24)
Number of reflections, test set	21679, 1119 (2384, 126)
Final R_{work}	0.2098 (0.2352)
Final R_{free}	0.2488 (0.2882)
Number of nonhydrogen atoms	3248
Protein residues	369
Ion (cadmium)	7
Water	112
r.m.s.d.	
Bonds (Å)	0.009
Angles (°)	1.111
Average *B*-factors (Å^2)	53.10
Ramachandran plot	
Most favoured (%)	97.25
Allowed (%)	2.75
Outliers (%)	0

asymmetric unit. Because the first five and the last 12 amino acids of the full length expression construct were not visible in the electron density maps, each Csm2 chain in the model comprises 123 amino acids. A total of seven cadmium ions and 112 molecules of water could be additionally fitted in the asymmetric unit of the crystals. Structure refinement statistics are summarized in Table 2.

The Csm2 fold

Each chain of Csm2 contained amino acids Gly6 to Ser128 of the protein. The Csm2 fold is formed by four α-helices (Fig. 1A). The two N-terminal (H1 and H2) helices and the C-terminal helix (H4) have a length of 18, 15 and 18 amino acids, respectively, whereas the third α-helix (H3) is larger in size and comprises 42 amino acids. Together, the helices form a paper clip like structure, with the N-terminal helices folded at one end, and the C-terminal helix folded at the other end of the larger α-helix.

Interestingly, all large helices from the three chains of Csm2 in the asymmetric unit of the structure are aligned with another large helix (either from the same or an adjacent asymmetric unit) in anti-parallel manner, in the form of dimers. The interaction between the helices follows loosely the pattern expected in coiled-coiled structures, and is mediated by leucine-zipper like patterns with hydrophobic residues present at regular positions of the helices. This dimeric structure

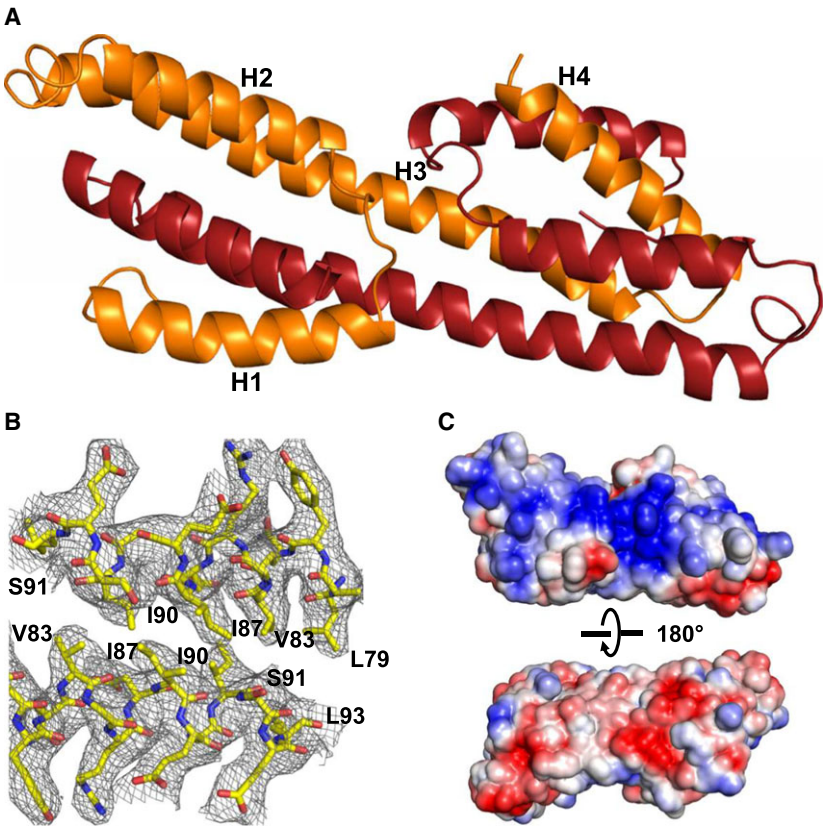


Fig. 1. The *T. maritima* Csm2 fold. (A) Structure of the Csm2 dimer. Helices of one of the Csm2 subunits are numbered H1 to H4 (orange). (B) Detail of the hydrophobic core of the Csm2 dimer showing interactions between residues of the larger helices (H3). (C) Electrostatic potential molecular surface representation of the Csm2 dimer in the same orientation (above) and rotated 180° (below) with respect to (A).

is further stabilized by interactions of the smaller N-terminal and C-terminal helices. Most notably, the N-terminal helix H1 is inserted into a gap between helix H3 and H4 of an adjacent subunit, forming exclusive interactions between these helices and, altogether, a helical packing with five α-helices at both ends of the dimer.

The interaction between the subunits in the dimer is mediated manly by hydrophobic residues, although numerous hydrogen bridges and salt bridges are also detectable at the interface. The total accessible surface area buried by the dimer formation amounts to approximately 4000 $Å^2$. A detail of the electron density of this interaction surface is shown in Fig. 1B.

Biochemical analysis

All these structural observations indicate that Csm2 forms a dimeric structure both in our crystals and *in vitro*. Because of this possibility, our recombinant protein purifications were further analyzed through size exclusion chromatography and MS. The initial results from the size exclusion chromatography experiments indicated the presence of a multimeric form of Csm2 in addition to the monomeric form of the protein (Fig. 2A), and this was confirmed by SDS/PAGE (Fig. 2B), MS (Fig. 2C,D) and a second round of size exclusion chromatography (Fig. 2E,F).

Deconvolution of the MS spectra derived from size exclusion chromatography fractions yielded isotope-averaged molecular masses compatible with both the multimeric and monomeric forms in the first fraction (Fig. 2C) and mainly the monomeric form in the second fraction (Fig. 2D). From the charge state distributions in the MS spectra, it was possible to accurately determine the masses of 16 752.13 ± 4.52 Da and 32 328.25 ± 47.60 Da for the monomeric and multimeric forms, respectively (Fig. 2C,D), thus confirming the dimerization of Csm2 in LC-MS experiments. Interestingly, both the monomeric and multimeric fractions of Csm2 remained intact in a further round of size exclusion chromatography experiments (Fig. 2E, F). Together, these results provide evidence that *T. maritima* Csm2 forms dimers *in vitro* and very probably also *in vivo*. These dimers are likely of great importance to the formation of the Csm crRNP complex and consequently to its function in oligonucleotide target recognition and degradation.

Implications for Csm crRNP complex formation

Recent electron microscopy and crystal structures have revealed in great detail the structure of the type I and type IIIB crRNP complexes [17,20,22–27,30,33,34]. Two recent electron microscopy structures solved, albeit at lower resolution, the multimer composition of type IIIA complexes from two different prokaryotic species, *S. solfataricus* [28] and *T. thermophilus* [18]. Here, we provide a high resolution structure of the *T. maritima* type III-A crRNP complex member Csm2 and demonstrate that it forms dimers *in vitro*. Because *T. maritima* Csm2 is a distant homologue of *S. solfataricus* SSo1424 (18% sequence identity) and *T. thermophilus* Csm2 (22% sequence identity), we used the structure solved here to obtain more information about the structural assembly and function of Csm2 in type IIIA crRNP complexes.

CLUSTALW analysis (Fig. 3A) shows that, despite low sequence identity, the three proteins could be functional and structural homologues because the pattern of important residues forming the hydrophobic core of the presumed α-helical bundles (underlined residues in Fig. 3A) is conserved. Interestingly, in this analysis, some of the loops connecting the helices showed a higher than average degree of conservation, indicating that they are possibly of high functional relevance. The recently obtained electron microscopy structures of the potential homologues of *S. solfataricus* and *T. thermophilus* reveal a multimeric complex of proteins, which are formed by an interwoven architecture of two protein complexes capped at both ends by additional proteins. The interwoven architecture of the complex is considered to be assembled mainly by the subunits of Sso 1424 and Sso1426 (in the case of *S. solfataricus*) and Csm2 and Csm3 (in the case of *T. thermophilus*) [18,28].

The whole complex is structurally similar to the structures of the Cascade and Cmr crRNP complexes, which have been solved at high resolution by protein crystallography [24–27]. In these structures, the interwoven architecture is assembled by two major helical protein filaments that form the 'backbone' (mainly involved in crRNA binding) and the 'belly' (which binds target oligonucleotides in the Cascade and Cmr complex) and are formed by Cas7 and Cse2 in Cascade [24,26] and Cmr4 and Cmr5 in Cmr crRNP [27] complexes, respectively.

We could not confirm structural similarity between Csm2 and Cas proteins from crRNP complexes such as the Cse2, Cmr5 or the related Csa5 protein [35]. Although these proteins have an α-helical fold, they do not contain a long central α-helix, and their helices are ordered in a different fold. It is still possible, however, that *T. maritima* Csm2, *S. solfataricus* SSo1424 and *T. termophilus* Csm2 function analogously as Cse2 and Cmr5 proteins of the Cascade and the Cmr crRNP

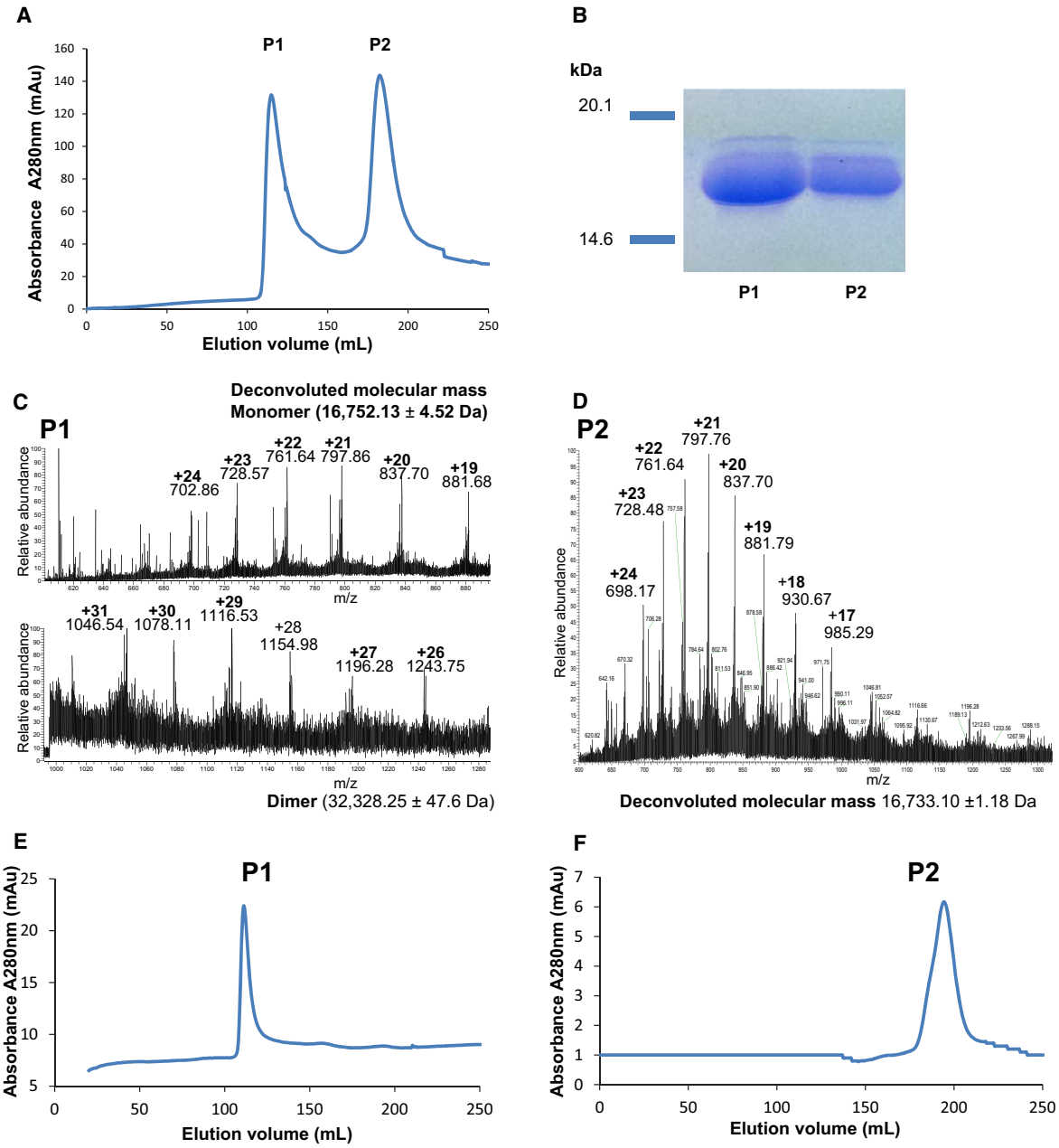


Fig. 2. Biochemical characterization of Csm2 dimerization. (A) Size exclusion chromatogram showing two Csm2 fractions. P1 (first fraction) represents a multimeric form of Csm2 and P2 (second fraction) represents a monomeric form. (B) SDS/PAGE of the corresponding fractions from size exclusion chromatography containing purified Csm2. (C) LC-MS analysis of the first fraction (P1). Deconvolution allowed identification of monomeric (above) and dimeric (below) Csm2. (D) LC-MS analysis of the second fraction (P2). Deconvolution allowed identification of monomeric Csm2. (E) Additional size exclusion chromatogram of the first fraction, demonstrating the intact multimer. (F) Additional size exclusion chromatogram of the second fraction demonstrating the presence of monomers.

complexes to bind target oligonucleotides [24,26,27]. Coincidentally, the Csm2 dimer shows a positive electrostatic potential on one side and a negative electrostatic potential on the other side (Fig. 1C), indicating that the first side could be involved in binding target oligonucleotides. In this case, the similarity of these subunits from type IIIA complexes with the subunits from type I and type IIIB complexes would be functional rather than structural, as postulated by Reeks *et al.* [35].

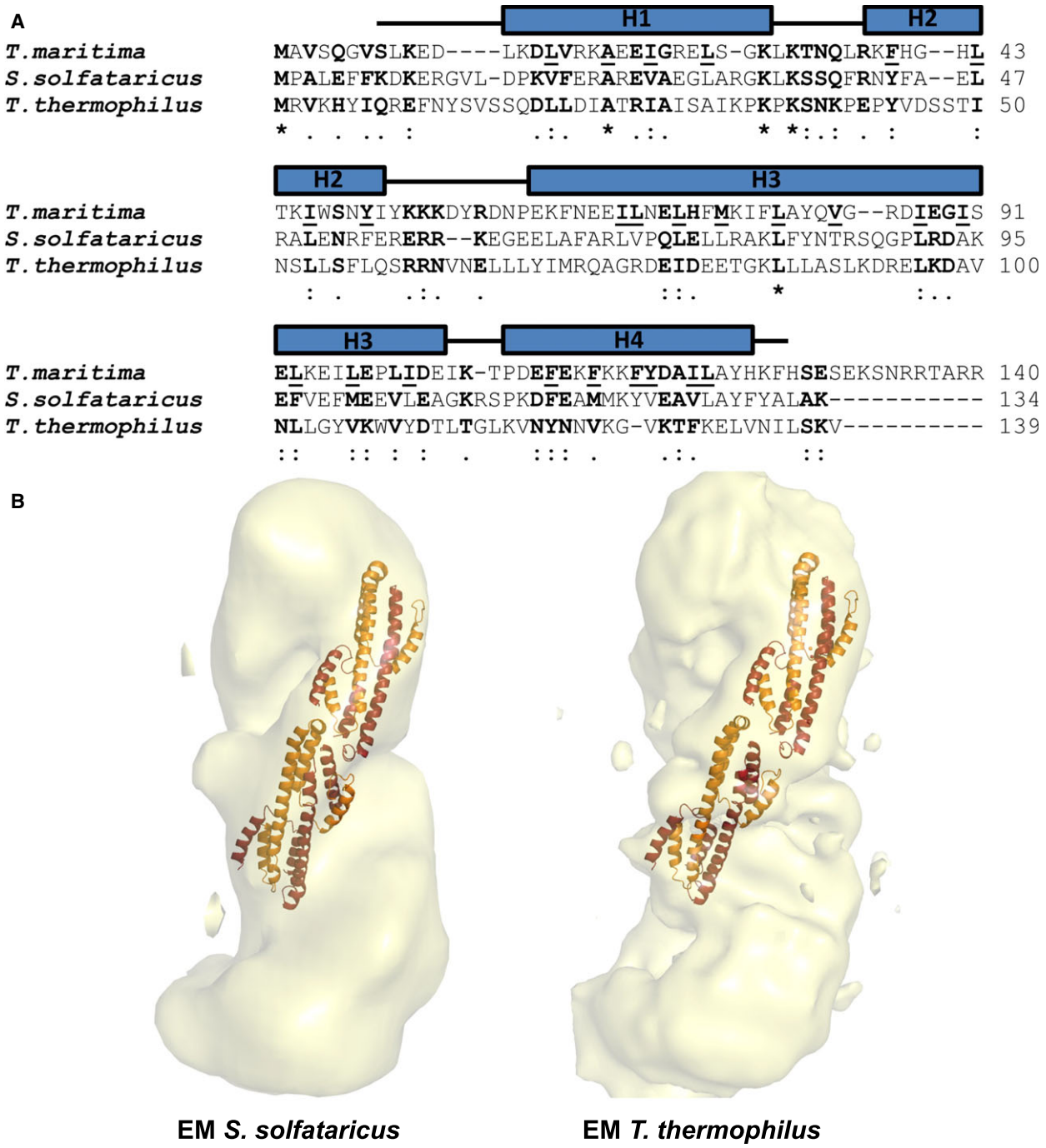


Fig. 3. Comparison of the Csm2 sequence and structure. (A) Multiple alignment of presumed Csm2 homologues from the three indicated prokaryotic species (*T. maritima* Csm2, *S. solfataricus* SSo1424 and *T. thermophilus* Csm2) indicating structural conservation. Similar amino acids are depicted in bold, where an asterisk (*) indicates complete conservation, a colon (:) indicates conservation between groups with strongly similar properties and a dot (.) indicates conservation between groups with weakly similar properties. Important amino acids forming the hydrophobic core of *T. maritima* Csm2 are underlined. (B) Fitting of *T. maritima* Csm2 crystal structures into the electron microscopy electron density maps of the of the crRNP complexes of *S. solfataricus* (left) and *T. thermophilus* (right).

Aiming to elaborate further on this hypothesis, we fitted our structure into the electron density maps of *S. solfataricus* (EMD-2420, 30 Å resolution) [28] and *T. termophilus* (EMD-6122, 17 Å resolution) [18]. As shown in Fig. 3B, up to two Csm2 dimers could be tentatively fitted into the electron density maps of

these two structures (Fig. 3B). The two Csm2 dimers were fitted as Csm2 tetramers to the region of the interwoven part opposite to the protein complex, which is considered to be formed by *S. solfataricus* SSo1426 or *T. termophilus* Csm2 [18,28]. This model indicates once more that Csm2 could be a structural and functional homologue of *S. solfataricus* SSo1424 and *T. termophilus* Csm2. Consistent with recent results indicating that, in Cmr complexes, the number of subunits of belly filaments can vary [27,30] and, in Csm complexes, the number of backbone subunits can vary to adapt to the variable sizes of crRNA [36], both dimer and tetramer formations of Csm2 would be compatible with a variable assembly of Csm RNPs in response to variable crRNA sizes.

As the Csm2 surface on the side with positive electrostatic potential forms a channel between the small α-helices, and the sequence of the loops forming this channel (located between helices H1 and H2) shows a higher than average degree of conservation in all three species, it is tempting to speculate that this channel is of high functional importance for example participating in the binding of target oligonucleotide bases. Additionally these loops, together with the N-terminal helices and contrasting with the large α-helices, have a higher than average B-factor distribution, indicating that they could be involved in ligand binding, as intrinsic flexibility is thought to be important in DNA binding proteins [37, 38]. However, as fitting of crystal structures to electron microscopy electron density maps is not trivial, these electron density maps are derived from different species and furthermore the number of Csm2 homologues detected in mass spectrometry data in crRNP complexes from these different species corresponds most probably to 3 subunits [18, 28], this evidence must be regarded as preliminary. More data from the Type IIIA RNA/Protein complexes will be needed for the confirmation of these hypotheses.

Conclusions

In conclusion, we were able to solve the structure of *T. maritima* Csm2, which is a major protein of the Csm crRNP complexes and is probably involved in assembling and binding target oligonucleotides to the crRNP complex. The results obtained indicate that Csm2 forms dimers, and thus could be a functional but not structural analogue of the Cascade protein Cse2 and the Cmr protein Cmr5. The dimeric structure of Csm2 described in the present study is therefore of great importance for better understanding the function of Csm crRNP complexes in type IIIA CRISPR/Cas systems, as well as the similarities and diversities between the different CRISPR/Cas systems.

Materials and methods

Protein expression, purification and crystallization

Full-length *T. maritima* Csm2 (GenBank entry AKE29563.1) was amplified via PCR with primer overhangs (Exxtend, Paulínia, Brazil) and cloned into the pQtev expression vector (Protein Structure Factory, Berlin, Germany). Csm2 was expressed in *E. coli* strain BL21(DE3) after induction with 1 mM IPTG. Cells were lysed in buffer containing 50 mM Tris-HCl (pH 7.5), 400 mM NaCl, 10 mM imidazole, 1 mM PMSF and 5 mM β-mercaptoethanol in an M-110L high pressure homogenizer (Microfluidics, Westwood, MA, USA). Protein was purified by His-tag affinity chromatography, the His-tag cleaved by TEV protease [39] and the protein further purified using size exclusion chromatography on a HiLoad 26/600 Superdex 75 prep grade column (GE Healthcare, Little Chalfont, UK; nominal volume of 320 mL) in buffer containing 25 mM Tris-HCl (pH 7.5), 100 mM NaCl, 3 mM DTT and 10% glycerol. After concentration, protein was crystallized by the hanging drop methods in crystallization conditions containing 100 mM cadmium chloride, 21% poly(ethylene glycol) 400 and 100 mM sodium acetate buffer (pH 4.6).

Data collection and processing

Crystals were measured at the MX-1 and MX-2 beamline of the Brazilian Synchrotron Light Laboratory (LNLS, Campinas, Brazil) [40,41] after cryo-cooling in liquid nitrogen using crystallization buffer supplemented with 12% glycerol. Datasets were measured at a wavelength of 1.458 Å. Crystallographic data were processed with XDS [42]. Crystal structure was solved by Cd-SAD [31] using PHASER [43] within the Autosol module of PHENIX and refined using CNS [44,45] and PHENIX.REFINE [46]. Electron density maps were inspected and the structural model built using COOT [47]. The electrostatic potential of the protein was calculated with APBS [48]. Images were drawn and rendered with PYMOL (http://www.schrodinger.com/pymol/). Surface area calculations were performed with PDBEPISA [49].

MS

The molecular mass of native Csm2 (monomeric and multimeric forms, derived from size exclusion chromatography) was confirmed by MS using an LTQ-Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, Waltham, MA, USA). MS spectra were acquired in the Orbitrap analyzer at 100 000 resolution (m/z 400). Nanoflow liquid chromatography was carried out on an Easy nLC nanoHPLC (Thermo

Fisher Scientific) coupled to the mass spectrometer. Proteins (5 μg) were loaded onto the column with buffer A (0.1% formic acid) and eluted with a 40 min gradient from 0% to 85% buffer B (acetonitrile, 0.1% formic acid). Protein mass spectra were visualized with the Qual Browser™ module, using XCALIBUR, version 3.0 (Thermo Fisher Scientific) and charge state deconvolution was undertaken after averaging 50–200 MS scans for each LC-MS run.

Acknowledgements

This work was supported by grant 11/50963-4 from FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo, São Paulo Research Foundation), as well as by grants 480411/2011-5 and 448833/2014-0 from CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, National Council for Scientific and Technological Development, Brazil). The authors also thank LNLS and LNLS beamline staff for beamline access and Dr Solange M. T. Serrano from the Instituto Butantan (São Paulo, Brazil) for help with the MS measurements.

Author contributions

GG cloned, purified and crystallized the protein, and also refined the structure and helped produce the illustrations. GA contributed to the cloning and purification of the protein, the writing of the manuscript, and the LC-MS experiments. GR contributed to the improvement of purification steps. AZ performed the LC-MS experiments. MAM contributed to the design of the study. CBC contributed to the design of the study, and reviewed the manuscript. MW designed the study and solved the structure. GG, MAM and MW wrote the manuscript. All authors approved the final manuscript submitted for publication.

References

1 Mojica FJ, Diez-Villasenor C, Garcia-Martinez J & Soria E (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* **60**, 174–182.
2 Bolotin A, Quinquis B, Sorokin A & Ehrlich SD (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551–2561.
3 Pourcel C, Salvignol G & Vergnaud G (2005) CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653–663.
4 Makarova KS, Grishin NV, Shabalina SA, Wolf YI & Koonin EV (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct* **1**, 7.
5 Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA & Horvath P (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712.
6 Marraffini LA & Sontheimer EJ (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**, 1843–1845.
7 Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV & van der Oost J (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960–964.
8 Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, Terns RM & Terns MP (2009) RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**, 945–956.
9 Wiedenheft B, Zhou K, Jinek M, Coyle SM, Ma W & Doudna JA (2009) Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. *Structure* **17**, 904–912.
10 Marraffini LA & Sontheimer EJ (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet* **11**, 181–190.
11 Grissa I, Vergnaud G & Pourcel C (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**, 172.
12 Makarova KS, Aravind L, Wolf YI & Koonin EV (2011) Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol Direct* **6**, 38.
13 Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF *et al.* (2011) Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* **9**, 467–477.
14 Garneau JE, Dupuis ME, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadan AH & Moineau S (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67–71.
15 Sinkunas T, Gasiunas G, Fremaux C, Barrangou R, Horvath P & Siksnys V (2011) Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J* **30**, 1335–1342.
16 Gasiunas G, Barrangou R, Horvath P & Siksnys V (2012) Cas9-crRNA ribonucleoprotein complex

mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc Natl Acad Sci USA* **109**, E2579–2586.
17 Staals RH, Agari Y, Maki-Yonekura S, Zhu Y, Taylor DW, van Duijn E, Barendregt A, Vlot M, Koehorst JJ, Sakamoto K *et al*. (2013) Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of *Thermus thermophilus*. *Mol Cell* **52**, 135–145.
18 Staals RH, Zhu Y, Taylor DW, Kornfeld JE, Sharma K, Barendregt A, Koehorst JJ, Vlot M, Neupane N, Varossieau K *et al*. (2014) RNA targeting by the type III-A CRISPR-Cas Csm complex of *Thermus thermophilus*. *Mol Cell* **56**, 518–530.
19 Samai P, Pyenson N, Jiang W, Goldberg GW, Hatoum-Aslan A & Marraffini LA (2015) Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. *Cell* **161**, 1164–1174.
20 Taylor DW, Zhu Y, Staals RH, Kornfeld JE, Shinkai A, van der Oost J, Nogales E & Doudna JA (2015) Structural biology. Structures of the CRISPR-Cmr complex reveal mode of RNA target positioning. *Science* **348**, 581–585.
21 Hale CR, Majumdar S, Elmore J, Pfister N, Compton M, Olson S, Resch AM, Glover CV III, Graveley BR, Terns RM *et al*. (2012) Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Mol Cell* **45**, 292–302.
22 Jore MM, Lundgren M, van Duijn E, Bultema JB, Westra ER, Waghmare SP, Wiedenheft B, Pul U, Wurm R, Wagner R *et al*. (2011) Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat Struct Mol Biol* **18**, 529–536.
23 Wiedenheft B, Lander GC, Zhou K, Jore MM, Brouns SJ, van der Oost J, Doudna JA & Nogales E (2011) Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature* **477**, 486–489.
24 Jackson RN, Golden SM, van Erp PB, Carter J, Westra ER, Brouns SJ, van der Oost J, Terwilliger TC, Read RJ & Wiedenheft B (2014) Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from *Escherichia coli*. *Science* **345**, 1473–1479.
25 Zhao H, Sheng G, Wang J, Wang M, Bunkoczi G, Gong W, Wei Z & Wang Y (2014) Crystal structure of the RNA-guided immune surveillance Cascade complex in *Escherichia coli*. *Nature* **515**, 147–150.
26 Mulepati S, Heroux A & Bailey S (2014) Structural biology. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. *Science* **345**, 1479–1484.
27 Osawa T, Inanaga H, Sato C & Numata T (2015) Crystal structure of the CRISPR-Cas RNA silencing Cmr complex bound to a target analog. *Mol Cell* **58**, 418–430.
28 Rouillon C, Zhou M, Zhang J, Politis A, Beilsten-Edmands V, Cannone G, Graham S, Robinson CV, Spagnolo L & White MF (2013) Structure of the CRISPR interference complex CSM reveals key similarities with cascade. *Mol Cell* **52**, 124–134.
29 Zhang J, Rouillon C, Kerou M, Reeks J, Brugger K, Graham S, Reimann J, Cannone G, Liu H, Albers SV *et al*. (2012) Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. *Mol Cell* **45**, 303–313.
30 Spilman M, Cocozaki A, Hale C, Shao Y, Ramia N, Terns R, Terns M, Li H & Stagg S (2013) Structure of an RNA silencing complex of the CRISPR-Cas immune system. *Mol Cell* **52**, 146–152.
31 Gallo G, Augusto G, Rangel G, Zelanis A, Mori MA, Barbosa Campos C & Wurtele M (2015) Purification, crystallization, crystallographic analysis and phasing of the CRISPR-associated protein Csm2 from *Thermotoga maritima*. *Acta Crystallogr F Struct Biol Commun* **71**, 1223–1227.
32 Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW *et al*. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D* **66**, 213–221.
33 Lintner NG, Kerou M, Brumfield SK, Graham S, Liu H, Naismith JH, Sdano M, Peng N, She Q, Copie V *et al*. (2011) Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). *J Biol Chem* **286**, 21643–21656.
34 Mulepati S & Bailey S (2011) Structural and biochemical analysis of nuclease domain of clustered regularly interspaced short palindromic repeat (CRISPR)-associated protein 3 (Cas3). *J Biol Chem* **286**, 31896–31903.
35 Reeks J, Naismith JH & White MF (2013) CRISPR interference: a structural perspective. *Biochem J* **453**, 155–166.
36 Hatoum-Aslan A, Samai P, Maniv I, Jiang W & Marraffini LA (2013) A ruler protein in a complex for antiviral defense determines the length of small interfering CRISPR RNAs. *J Biol Chem* **288**, 27888–27897.
37 Sunami T & Kono H (2013) Local conformational changes in the DNA interfaces of proteins. *PLoS One* **8**, e56080.
38 Craveur P, Joseph AP, Esque J, Narwani TJ, Noel F, Shinada N, Goguet M, Leonard S, Poulain P, Bertrand O *et al*. (2015) Protein flexibility in the light of structural alphabets. *Front Mol Biosci* **2**, 20.
39 Blommel PG & Fox BG (2007) A combined approach to improving large-scale production of tobacco etch virus protease. *Protein Expr Purif* **55**, 53–68.
40 Polikarpov I, Perles LA, de Oliveira RT, Oliva G, Castellano EE, Garratt RC & Craievich A (1998)

Set-up and experimental parameters of the protein crystallography beamline at the Brazilian National Synchrotron Laboratory. *J Synchrotron Radiat* **5**, 72–76.

41 Guimaraes BG, Sanfelici L, Neuenschwander RT, Rodrigues F, Grizolli WC, Raulik MA, Piton JR, Meyer BC, Nascimento AS & Polikarpov I (2009) The MX2 macromolecular crystallography beamline: a wiggler X-ray source at the LNLS. *J Synchrotron Radiat* **16**, 69–75.

42 Kabsch W (2010) Xds. *Acta Crystallogr D* **66**, 125–132.

43 McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC & Read RJ (2007) Phaser crystallographic software. *J Appl Crystallogr* **40**, 658–674.

44 Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS *et al.* (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr D* **54**, 905–921.

45 Brunger AT (2007) Version 1.2 of the Crystallography and NMR system. *Nat Protoc* **2**, 2728–2733.

46 Afonine PV, Grosse-Kunstleve RW, Echols N, Headd JJ, Moriarty NW, Mustyakimov M, Terwilliger TC, Urzhumtsev A, Zwart PH & Adams PD (2012) Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr D* **68**, 352–367.

47 Emsley P, Lohkamp B, Scott WG & Cowtan K (2010) Features and development of Coot. *Acta Crystallogr D* **66**, 486–501.

48 Baker NA, Sept D, Joseph S, Holst MJ & McCammon JA (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci USA* **98**, 10037–10041.

49 Krissinel E & Henrick K (2007) Inference of macromolecular assemblies from crystalline state. *J Mol Biol* **372**, 774–797.