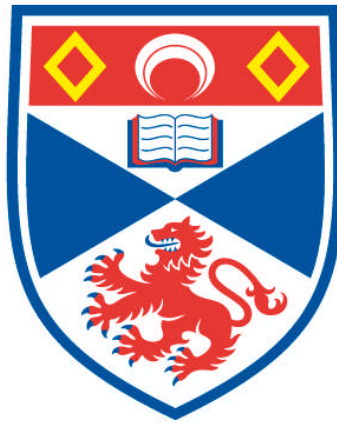


**CHARACTERISATION OF PROTEINS INVOLVED IN
CRISPR-MEDIATED ANTIVIRAL DEFENCE IN
SULFOLOBUS SOLFATARICUS**

Melina Louiza Kerou

**A Thesis Submitted for the Degree of PhD
at the
University of St Andrews**



2012

**Full metadata for this item is available in
Research@StAndrews:FullText
at:**

<http://research-repository.st-andrews.ac.uk/>

Please use this identifier to cite or link to this item:

<http://hdl.handle.net/10023/3088>

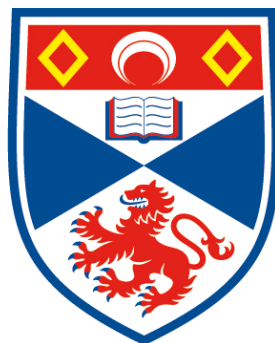
This item is protected by original copyright

Characterisation of proteins involved in
CRISPR-mediated antiviral defence
in *Sulfolobus solfataricus*

Melina Louiza Kerou

A Thesis Submitted for the Degree of
Doctor of Philosophy

March 2012



University of
St Andrews

Table of contents

Figures and tables.....	vii
Abbreviations.....	xii
Declaration.....	xv
Abstract.....	xvii

Acknowledgements

xviii

Introduction.....	1
1.1 Defence mechanisms in prokaryotes.....	1
1.1.1 Restriction-modification systems.....	3
1.1.2 Abortive infection.....	4
1.1.3 Dissemination, evolution and additional roles of defence systems.....	6
1.2 Discovery and characterisation of the CRISPR arrays in prokaryotic genomes.....	8
1.3 CRISPR-associated protein families and current classification of the CRISPR/Cas system.....	13
1.4 The three stages of the CRISPR/Cas mode of action.....	17
1.4.1 Stage I: Spacer selection and integration into CRISPR arrays.....	17
1.4.2 Stage II: CRISPR expression and biogenesis of crRNAs.....	21
1.4.2.1 Regulation of CRISPR transcription.....	22
1.4.2.2 CRISPR transcript processing.....	23
1.4.3 Stage III: Recognition of invader sequences and target interference.....	29
1.4.4 Protospacer selection, self/non-self discrimination and autoimmunity issues.....	34
1.5 Evolution, mobility and distribution of the CRISPR/Cas system.....	37
1.6 CRISPR/Cas and the eukaryotic RNAi.....	38
1.7 Applications and alternative roles for CRISPR/Cas.....	40

1.8 CRISPR/Cas systems of <i>Sulfolobus solfataricus</i>	41
Materials and Methods	45
2.1 Cloning procedures	45
2.1.1 Cloning and vectors	45
2.1.2 Site-directed mutagenesis	46
2.2 Protein expression and purification	46
2.2.1 Expression of recombinant proteins	46
2.2.2 Purification of recombinant proteins	47
2.3 Crystallization screening and optimisation	49
2.4 Immuno - blot.....	50
2.5 Protein Interactions	50
2.5.1 Analytical size exclusion chromatography	50
2.5.2 Determination of protein interactions using magnetic precharged nickel particles	51
2.6 Generation of nucleic acid substrates and markers	52
2.6.1 Purification of oligonucleotides	52
2.6.2 Assembly and purification of double-strand substrates.....	54
2.6.3 CRISPR locus constructs	55
2.6.4 T7 RNA polymerase-mediated in vitro transcription	55
2.6.5 Sanger DNA sequencing	56
2.6.6 RNA alkaline hydrolysis ladder.....	56
2.7 <i>Sulfolobus solfataricus</i> in vitro transcription.....	57
2.8 Extraction of RNA from purified native aCASCADE	57
2.9 Nucleic acid binding and catalytic assays	58
2.9.1 Helicase assays	58
2.9.2 Endonuclease assays	58
2.9.3 ATP hydrolysis reaction	59
2.9.4 Electrophoretic Mobility Shift Assay (EMSA).....	59
2.9.5 Strand annealing and strand exchange assays.....	60

2.9.6 R-loop formation assay	60
The CMR complex from <i>Sulfolobus solfataricus</i> : native isolation and recombinant components	63
3.1 Introduction	63
3.1.1 The Repeat-Associated Mysterious Proteins (RAMPs)	63
3.1.2 The CRISPR-RAMP module (CMR)	65
3.2 Expression and purification of recombinant Cmr7	69
3.3 Crystallographic study of Cmr7	70
3.3.1 Crystallisation and structure solution of Cmr7	70
3.3.2 Structure of Cmr7	71
3.5 Expression and purification of recombinant Cmr1	74
3.6 Protein interactions between recombinant CMR components	75
3.7 Nucleic acid binding by recombinant CMR proteins	78
3.8 Isolation of the native CMR complex from <i>Sulfolobus solfataricus</i>	80
3.8.1 Antibody assisted purification of the SsoCMR complex	80
3.8.2 Identification of the CMR complex components by mass spectrometry	84
3.9 Initial functional characterisation of the native SsoCMR complex	87
3.9.1 The SsoCMR complex does not bind ssDNA or ssRNA	87
3.9.2 The SsoCMR complex does not exhibit nuclease activity against CRISPR RNA	89
3.9.3 The SsoCMR complex does not exhibit polymerase activity	90
3.8 Discussion	92
Purification and characterisation of Csa2-Cas5a: An archaeal CASCADE-like complex for CRISPR-mediated viral defence	98
4.1 Introduction	98
4.1.1 Biochemical and structural characterisation of the <i>E. coli</i> CASCADE	98
4.1.2 An archaeal orthologue of CASCADE	101
4.2 Site-directed mutagenesis of Csa2	103
4.3 Expression and purification of recombinant wild-type and mutant Csa2 and the Csa2-Cas5a complex	103

4.4 Investigation of the native Csa2-Cas5a complex from <i>Sulfolobus solfataricus</i> P2 and its accessory proteins	107
4.4.1. Nucleic acid content of the native aCASCADE	109
4.5 Size determination and stoichiometry of the native and recombinant Csa2-Cas5a complex.....	111
4.6 In vitro protein interactions of the recombinant Csa2-Cas5a complex.	113
4.7 Investigating the properties of the leader sequence of CRISPR locus A	115
4.9 The recombinant Csa2-Cas5a complex binds crRNA and forms ternary complexes with target DNA	122
4.9.1 Substrate analysis of crRNA	123
4.9.2 Csa2 is the main crRNA binding subunit of the Csa2-Cas5a complex	125
4.9.3 The crRNA - loaded Csa2-Cas5a complex recognises and binds target DNA	126
4.9.4 The protospacer adjacent motif (PAM) is not required for target DNA recognition by the recombinant Csa2-Cas5a complex.....	129
4.10 Structural studies and discussion	131
4.10.1 The structure of Csa2.....	131
4.10.2 Arrangement of the native aCASCADE and mechanistic implications	135
4.10.3 Emerging model for CRISPR-mediated interference in Archaea.....	137
Initial biochemical characterisation of Cas3' from <i>S. solfataricus</i> : a predicted CRISPR-associated helicase.....	142
5.1 Introduction	142
5.1.2 The DExD/H-box families of RNA-remodeling proteins	144
5.1.3 The CRISPR-associated putative DExH-box helicase Cas3	147
5.2 Cas3' in <i>Sulfolobus solfataricus</i>	152
5.3 Expression and purification of SsoCas3'	153
5.4 Site-directed mutagenesis of SsoCas3' (Sso1440)	154
5.5 ATPase activity of SsoCas3'	156
5.6 Helicase activity and substrate preference of SsoCas3'	159
5.6.1 SsoCas3' is not able to process long duplex regions	165
5.7 Nucleic acid binding by SsoCas3'.....	166
5.8 Strand annealing and strand exchange activity of SsoCas3'	167

5.8.1 Initial attempt to investigate R-loop formation by SsoCas3'	171
5.9 Discussion and future work.....	172
Conclusions and future work	178
References.....	183
APPENDIX I.....	203
CRISPR locus constructs	203
APPENDIX II	204
Multiple sequence alignments.....	204
A.2.1 Conserved motifs of Cmr2 family members	204
A.2.2 Multiple sequence alignment of Csa2 orthologues	208
A.2.3 Multiple sequence alignment of Cas6 orthologues	210

Figures and tables

Figure 1.1	Classification of phage infection types and prokaryotic resistance mechanisms in respect to their effect on the fitness of the host and the phage.	2
Figure 1.2	The general course of a phage or viral infection	3
Figure 1.3	Outline of toxin-antitoxin systems	5
Figure 1.4	Characteristic behavior of selfish genetic elements	7
Figure 1.5	Graphic representation of a CRISPR locus and the adjacent cas gene operon in a prokaryotic genome	10
Figure 1.6	Outline of the CRISPR/Cas mode of action	12
Figure 1.7	Outline of the main types and subtypes of the CRISPR/Cas systems and their phylogenetic relations	14
Figure 1.8	Crystal structure of Cas1	15
Figure 1.9	Crystal structure of Cas2	16
Figure 1.10	Structure of putative transcriptional regulator Csa3 from <i>S. solfataricus</i>	23
Figure 1.11	Outline of the second stage of CRISPR functioning	23
Figure 1.12	Structures of processing endonucleases Cse3 and Csy4	25
Figure 1.13	Crustal structure of PfuCas6	26
Figure 1.14	Model for CRISPR RNA processing in type II systems	28
Figure 1.15	Gene organisation of Cas operons of studied organisms	32
Figure 1.16	Basepairing between crRNA and protospacer upon target recognition in <i>E. coli</i>	33
Figure 1.17	Model for target recognition by CASCADE	34
Figure 1.18	Orientation of protospacers in regard to their PAM	35
Figure 1.19	Model for discriminating between self and non-self DNA during CRISPR target recognition	36
Figure 1.20	General pathway and key proteins for RNA interference	39
Figure 1.21	Cas genes in <i>S. solfataricus</i> P2	42
Figure 2.1	Amplified fragments of the <i>S. solfataricus</i> CRISPR locus A	55
Figure 2.2	Mechanism of RNA hydrolysis under alkaline conditions	56
Figure 3.1	Conserved RAMP superfamily motifs and individual RAMP families	64

Figure 3.2	Crystal structure of Tthb192 - CasE/Cse3 homologue	64
Figure 3.3	Crystal structures of Cmr5	67
Figure 3.4	Gene organisation of the type III systems in <i>S. solfataricus</i>	68
Figure 3.5	Domain organisation of SsoCmr2 (Sso1991), indicating the location of the conserved potential active sites on the sequence.	68
Figure 3.6	Purification of Cmr7	70
Figure 3.7	Crystals of Cmr7, grown under the conditions mentioned in the text	71
Figure 3.8	Structure of the Cmr7 monomer	72
Figure 3.9	Cmr7 dimer and conserved surfaces	73
Figure 3.10	Purification of Cmr4	74
Figure 3.11	Purification of Cmr1	75
Figure 3.12	Experimental setup for detecting protein-protein interactions using Ni-loaded magnetic beads	76
Figure 3.13	Interactions between recombinant Cmr subunits	77
Figure 3.14	Electrophoretic Mobility Shift Assays investigating the binding affinity of recombinant Cmr proteins to various nucleic acid substrates, carried out in collaboration with Paul Talbot	79
Figure 3.15	Effect of Cmr3 on Cmr1 RNA binding	80
Figure 3.16	First step of SsoCMR purification by affinity chromatography	81
Figure 3.17	Second step of SsoCMR purification by size exclusion chromatography.	81
Figure 3.18	Third step of SsoCMR purification by cation exchange chromatography	82
Figure 3.19	Fourth step of SsoCMR purification by anion exchange chromatography.	83
Figure 3.20	Optimisation of the SsoCMR purification procedure by Paul Talbot with a larger starting culture.	84
Figure 3.21	Binding of native SsoCMR to CRISPR ssDNA and RNA substrates.	88
Figure 3.22	Substrates used for assaying the nuclease activity of the SsoCmr complex	89
Figure 3.23	Nuclease assays of SsoCMR	90
Figure 3.24	Outline of in vitro transcription with the SsoRNA polymerase	91
Figure 3.25	Reverse transcriptase assays for SsoCMR	92
Figure 3.26	Mode of action of the PfuCmr complex	93

Figure 4.1	Structure of the <i>E. coli</i> CASCADE and the Csy complex from <i>P. aeruginosa</i>	100
Figure 4.2	Gene names and operon organisation of type I-A CRISPR/Cas in <i>S. solfataricus</i>	102
Figure 4.3	Purification of recombinant Csa2-Cas5a and Csa2 WT, H160A	105
Figure 4.4	ESI-TOF mass spectrometry of Csa2 WT and H160A	106
Figure 4.5	Isolation of native aCASCADE from <i>S. solfataricus</i>	107
Figure 4.6	Nucleic acid content of the aCASCADE	110
Figure 4.7	Analytical size-exclusion chromatography on native and recombinant Csa2-Cas5a complex	112
Figure 4.8	Experimental setup for <i>in vitro</i> protein interaction assay	114
Figure 4.9	Protein interactions of the recombinant Csa2-Cas5a complex	115
Figure 4.10	CRISPR locus A constructs	116
Figure 4.11	In vitro transcription of CRISPR locus with SsoRNAP	116
Figure 4.12	Mapping the transcription initiation site	118
Figure 4.13	Multiple sequence alignment of leader sequences of <i>S. islandicus</i> , <i>S. solfataricus</i> 98/2, <i>S. solfataricus</i> P2 CRISPR A and B and <i>Acidianus hospitalis</i>	118
Figure 4.14	Structure of PfuCas6 and model of SsoCas6	120
Figure 4.15	CRISPR transcript processing by SsoCas6	121
Figure 4.16	Binding of Csa2-Cas5a to crRNA	124
Figure 4.17	Comparative binding of aCASCADE individual subunits and the complex to crRNA	125
Figure 4.18	cr-RNA mediated binding of Csa2-Cas5a to DNA target	127
Figure 4.19	Absence of nuclease activity on DNA protospacer targets by aCASCADE	128
Figure 4.20	Effect of the Protospacer Adjacent Motif on crRNA -guided binding of DNA targets by the aCASCADE.	130
Figure 4.21	Structure of SsoCsa2	133
Figure 4.22	Electrostatic surface map of Csa2	134
Figure 4.23	Quaternary structural models of native aCASCADE	136
Figure 4.24	Emerging model for CRISPR interference in Archaea	140
Figure 5.1	Sequence and structural organization of the conserved motifs of SF1 and SF2 NTPases - translocases	143
Figure 5.2	Phylogenetic relationships and motif conservation between the SF2 families and between SF1 and SF2 families	144
Figure 5.3	Comparison of DExH-box and DEAD-box RNA helicases	146

Figure 5.4	Proposed mechanism of action for Cas3 in type I-E systems	150
Figure 5.5	Structure of the HD-domain of Cas3 from <i>T. thermophilus</i>	151
Figure 5.6	Domain arrangement of SsoCas3'	152
Figure 5.7	Structural model of SsoCas3', generated by Phyre2	153
Figure 5.8	Purification of recombinant SsoCas3'	154
Figure 5.9	ESI-TOF mass spectrometry of SsoCas3' WT and K46A	155
Figure 5.10	ATPase activity of WT and mutant SsoCas3'	158
Figure 5.11	Helicase activity of SsoCas3' on CRISPR substrates	161
Figure 5.12	Temperature dependance on the helicase activity of SsoCas3'	162
Figure 5.13	Helicase activity of SsoCas3' is dependent on protein concentration	162
Figure 5.14	Helicase activity of SsoCas3' on blunt RNA-DNA heteroduplexes	163
Figure 5.15	SsoCas3' is unable to unwind radiolabeled non-CRISPR related substrates	164
Figure 5.16	SsoCas3' is unable to unwind "crRNA-protospacer"-like substrates	165
Figure 5.17	SsoCas3' is unable to unwind long duplex substrates	166
Figure 5.18	Nucleic acid binding by SsoCas3'	167
Figure 5.19	Strand annealing activity of SsoCas3'	168
Figure 5.20	Strand exchange activity of SsoCas3'	170
Figure 5.21	Potential R-loop formation by SsoCas3'	172

Table 1.1	Taxonomic distribution of CRISPR-Cas systems in 706 analysed genomes	11
Table 1.2	CRISPR loci in <i>S. solfataricus</i> P2	42
Table 1.3	Characteristics of the Cas proteins in <i>S. solfataricus</i> P2	43
Table 2.1	Expression conditions of Cas proteins from <i>S. solfataricus</i>	47
Table 2.2	Oligonucleotides used in chapter 3	52
Table 2.3	Oligonucleotides used in chapter 4	53
Table 2.4	Oligonucleotides used in chapter 5	54
Table 3.1	The Cmr family B cluster in <i>S. solfataricus</i>	69
Table 3.2	Mass Spectrometry analysis of two different purifications of SsoCMR.	86
Table 4.1	Composition of the <i>E. coli</i> CASCADE	100
Table 4.2	Co-purifying Cas proteins in the partially purified native aCASCADE sample as identified by solution trypsin digestion followed by LC - MS/MS	108
Table 4.3	Synthetic oligonucleotides used in chapter 4	122
Table 4.4	Apparent Kd values for Csa2-Cas5a and the various RNA substrates	124
Table 5.1	Oligonucleotides used to generate substrates used in chapter 5	159
Table 5.2	Structures of ds substrates used for helicase assays	160

Abbreviations

3'	3 prime DNA end
5'	5 prime DNA end
a.a.	aminoacid
Abi	Abortive infection
[γ - ³² P] ATP	Adenosine triphosphate with a ³² -phosphate radioactive isotope in the gamma phosphate position
Afu	<i>Archaeoglobus fulgidus</i>
AMP-PNP	Adenosine 5'-(β , γ -imido) triphosphate
APS	Ammonium persulphate
ATP	Adenosine 5'-triphosphate
ATP- γ S	adenosin 5'-(O)-(3-thio)triphosphate
bp	base pair
BSA	Bovine serum albumin
Cas	CRISPR-associated
CMR	CRISPR module RAMP
COG	Cluster of orthologous groups
CRISPR	Clustered regularly interspersed short palindromic repeat
CV	Column volume
ds / ssDNA	Double / single-stranded deoxyribonucleic acid
DTT	1,4 – dithiothreitol
EDTA	Ethylenediaminetetraacid acid
EMSA	Electrophoretic mobility shift assay
ESI-TOF MS	Electrospray ionisation-time of flight mass spectrometry
GTP	guanosine 5'-triphosphate
HCl	Hydrochloric acid
HGT	Horizontal gene transfer
HMM	Hidden Markov Model
IPTG	Isopropyl-beta-D-thiogalactopyranoside

KD	Dissociation constant
(k)Da	(Kilo)Dalton
KOH	Potassium hydroxide
LB	Luria Bertani
LGT	Lateral gene transfer
MALDI-TOF	Matrix-assisted laser desorption/ ionisation-time of flight
MS	Mass spectrometry
MT	Methyltransferase
MW	Molecular weight
NA	Nucleic acid
nt	nucleotide
NTP	Nucleotide triphosphate
OD	Optical density
PBS	Phosphate buffered saline
Pfu	<i>Pyrococcus furiosus</i>
PNK	Polynucleotide kinase
RAMP	Repeat associated mysterious proteins
RE	Restriction endonuclease
RM	Restriction-modification
RNA	Ribonucleic acid
RNApol	RNA polymerase
RNP	Ribonucleoprotein
rpm	Revolutions per minute
RT	Room temperature
Sac	<i>Sulfolobus acidocaldarius</i>
SAXS	Small-angle X-ray scattering
SDS	Sodium dodecyl sulphate
SDS-PAGE	SDS-polyacrylamide gel electrophoresis
SF	Superfamily
Sso	<i>Sulfolobus solfataricus</i>
Sto	<i>Sulfolobus tokodaii</i>
TA	Toxin-antitoxin

TEM	Transmission electron microscopy
TBE	Tris-borate EDTA
TEMED	Tetramethylethylenediamine
TEV	Tobacco Etch virus
TPB	Tryptone-phosphate broth
UV	Ultraviolet
WT	Wild type

Declaration

1. Candidate's declarations:

I, Melina - Louiza Kerou, hereby certify that this thesis, which is approximately 64000 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in October 2007 and as a candidate for the degree of PhD in October 2008; the higher study for which this is a record was carried out in the University of St Andrews between 2007 and 2010.

Date signature of candidate

2. Supervisor's declaration:

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date signature of supervisor

3. Permission for electronic publication: *(to be signed by both candidate and supervisor)*

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the electronic publication of this thesis:

Access to printed copy but embargo of of electronic publication of thesis for a period of 1 year on the following ground(s):

publication would be commercially damaging to the researcher, or to the supervisor, or the University;

publication would preclude future publication;

Date signature of candidate signature of supervisor

Abstract

One of the most surprising realisations to emerge from metagenomics studies in the early '00s was that the population of viruses and phages in nature is about 10 times larger than the population of prokaryotic organisms. Thus, bacteria and archaea are under constant pressure to develop resistance methods against a population of viruses with extremely high turnover and evolution rates, in what has been described as an evolutionary "arms race". A novel, adaptive and heritable immune system encoded by prokaryotic genomes is the CRISPR/Cas system. Arrays of clustered regularly interspersed short palindromic repeats (CRISPR) are able to incorporate viral or plasmid sequences which are then used to inactivate the corresponding invader element via an RNA interference mechanism. A number of CRISPR-associated (Cas) protein families are responsible for the maintenance, expansion and function of the CRISPR loci. This system can be classified in a number of types and subtypes that differ widely in their gene composition and mode of action.

This thesis describes the biochemical characteristics of CRISPR-mediated defense in the crenarchaeon *Sulfolobus solfataricus*. The process of CRISPR loci transcription and their subsequent maturation into small guide crRNA units by the processing endonuclease of the system (Cas6) is investigated. After this step, different pathways and effector proteins are involved in the recognition and silencing of DNA or RNA exogenous nucleic acids. This thesis reports the identification and purification of a native multiprotein complex from *S. solfataricus* P2, the Cmr complex, a homologue of which has been found to recognise and cleave RNA targets in *P. furiosus*. The recognition and silencing of DNA targets in *E. coli* has been shown to involve a multiprotein complex termed CASCADE as well as Cas3, a putative helicase-HD nuclease. *S. solfataricus* encodes orthologues for the core proteins of this complex, and the formation and function of an archaeal CASCADE is investigated in this thesis.

Acknowledgements

First and foremost, I would like to thank my supervisor Professor Malcolm White for his scientific guidance throughout this project, his enormous patience and support, and for giving me the opportunity to participate in a very active scientific community based on international collaboration such as the Marie Curie networks. Secondly, i'd like to thank the members of the SSPF for their contributions and help on the CRISPR project and especially Dr Huanting Liu, who has been the most inspiring and amazing teacher i ever had. I am very grateful to Dr Sonja Albers and her group at the Max Planck Institute in Marburg for their hospitality and for sharing their expertise on the *Sulfolobus* genetic system. Thanks also to members of the Mass spectrometry group for analysing endless protein samples.

Naturally, more than half of what I learned these years came from sharing everyday lab life with a fantastic group of people. Thanks to all members of the White and Coote lab (past and present), and especially Dr Christophe Rouillon, Dr Sonia Paytubi and Dr Shirley Graham for their friendship and encouragement. Thanks to Dr Lester Carter, Dr Arif Sheikh and Dr Muse Oke for not being serious :) Enormous thanks to Professor Garry Taylor and Dr Margaret Taylor (and their cats!) for their encouragement, support and kindness all these years.

Finally, I want to thank my friends in St Andrews and Edinburgh for making these years an amazing experience. Without their support i would never have made it. Last but not least, my warmest thanks go to my parents for encouraging me to make this step and for always being there for me.

Chapter 1

Introduction

“Well, in our country,” said Alice, still panting a little, “you’d generally get to somewhere else - if you run very fast for a long time, as we’ve been doing.”

“A slow sort of country!” said the Queen. “Now, here, you see, it takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!”

-- Lewis Carroll, *Through the Looking-Glass*, 1871

1.1 Defence mechanisms in prokaryotes

This excerpt from the Red Queen’s race in Lewis Carroll’s novel *Through the Looking-Glass* illustrates one of the most influential hypotheses in evolutionary biology, termed the Red Queen’s Hypothesis. Originally proposed by Van Valen in 1973, the hypothesis states that in a dynamic evolutionary system, constant adaptation of all co-dependent species is necessary to maintain their fitness relative to one another. In a system comprising of species with a predator-prey relationship, a higher evolutionary rate exhibited by one species ensures the accumulation of fitness-increasing adaptations via natural selection. This provides the species in question with a competitive advantage over the slow-evolving species, which leads to a decrease in their fitness and can compromise their survival chances. Therefore, in order to maintain a relative fitness balance and ensure their survival, co-evolving species must exhibit comparably high evolutionary rates. In the prokaryotic kingdom, this hypothesis is especially accurate in describing virus-host dynamics and the evolution of molecular resistance and anti-resistance mechanisms (Stern and Sorek, 2010).

Resistance mechanisms are found at every stage of phage or virus infection (figures 1.1, 1.2) with variable outcomes on the phage and host survival (figure 1.1). Passive defence strategies mainly rely on adsorption resistance mechanisms, which prevent the attachment of a phage or virus to appropriate receptors on the prokaryotic cell surface (reviewed in Hyman and Abedon, 2010). A wide range of surface exposed molecules can act as virus receptors, including proteins, lipopolysaccharides, teichoic acids and capsules. One of the mechanisms which confer resistance is the production of layers of extracellular polymers which shield the receptor molecules. The

composition and efficiency of these polymers, as well as their mode of production, vary widely among bacteria, and phage-encoded counter-resistance has been detected in the form of enzymes which degrade certain polymers. A more effective mechanism is the loss or structural modification of a receptor molecule so that it will not be recognised by phages/viruses. Random mutations are usually the cause of the structural modifications, and sometimes the physiological receptor function might not be affected (Hyman and Abedon, 2010). The next line of defence operates at the post-adsorption level, in order to prevent the virus or phage genetic material from entering the host cytoplasm and taking over the host metabolism. The outcome of these systems can be positive for the prokaryotic cell and deleterious to the phage/virus. Prophage-encoded superinfection exclusion mechanisms (Sie), which prevent the injection of the phage DNA into the cytoplasm, are mostly common in Gram-negative bacteria (Mahony *et al.* 2008). Some of the most well-studied active resistance mechanisms also operate at this stage, including the restriction-modification system (RM) and the newly discovered CRISPR/Cas system (Stern and Sorek, 2010). Finally, in the case of a successful infection a number of systems lead to abortive infection (Abi), a term which describes the controlled “suicide” of the infected cell in order to completely prevent the release and spread of the new virus particles (Chopin *et al.* 2005, Hyman and Abedon, 2010, Stern and Sorek, 2010). The restriction-modification and abortive infection systems will be briefly presented and the rest of the chapter will focus on the novel adaptive prokaryotic immune system encoded by the CRISPR/Cas loci in the majority of bacteria and archaea.

	Phage lives	Phage dies
Bacterium lives	Adsorption resistance Phage-encounter blocks Receptor loss Lysogenic infection Absence of resistance Chronic infection Absence of resistance	Restriction Uptake blocks Restriction-modification CRISPR Superinfection exclusion Superinfection immunity Lysogeny blocks
Bacterium dies	Lytic infection Absence of resistance Reduced phage productivity Adsorption slowed Lysis delayed Burst size reduced Dissemination interference	Abortive infection Transcription blocks Translation blocks Replication blocks Maturation blocks Premature lysis Failure to lyse

Figure 1.1 Classification of phage infection types and prokaryotic resistance mechanisms in respect to their effect on the fitness of the host and the phage.

Adapted from Hyman and Abedon, 2010.

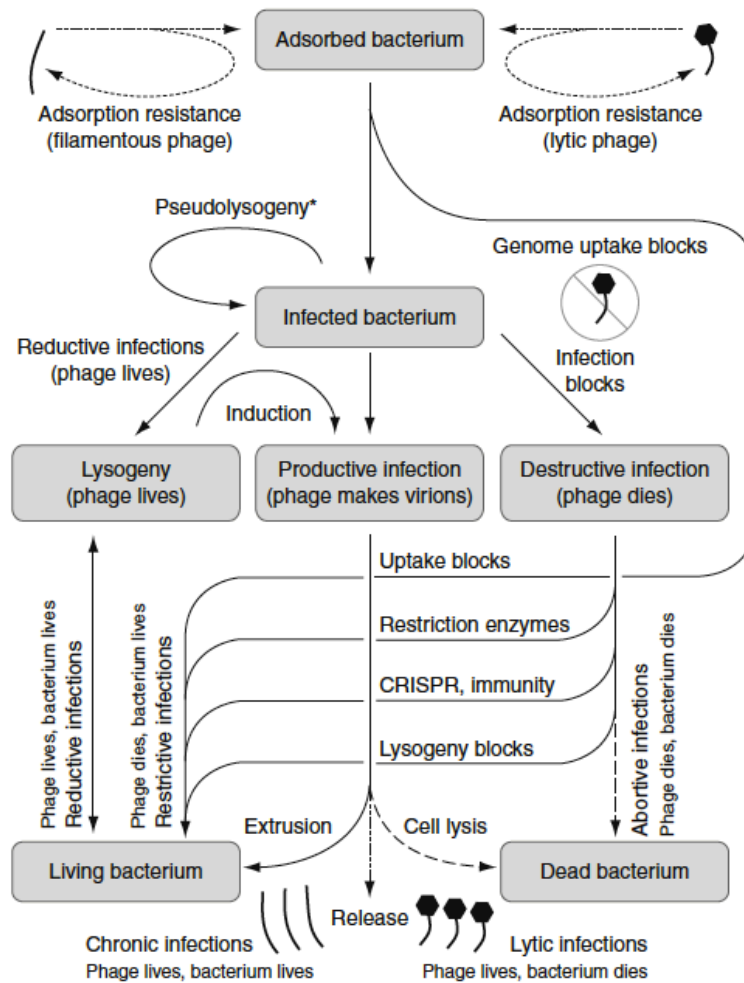


Figure 1.2: The general course of a phage or viral infection

The scheme includes the defence mechanisms likely to be encountered at each stage of the infection and all possible outcomes (adapted from Hyman and Abedon, 2010).

1.1.1 Restriction-modification systems

Restriction-modification (RM) systems are encoded by almost 90% of sequenced bacterial and archaeal genomes (Roberts *et al.* 2010), and can confer resistance to a wide variety of extrachromosomal elements (viruses, phages and plasmids, reviewed in Hyman and Abedon, 2010; Tock and Dryden, 2005). Two types of enzymes compose the core of this system: restriction endonucleases (REases), which perform sequence specific cleavage of foreign DNA, and methyltransferases (MTases), which protect the endogenous DNA from cleavage by modifying specific bases in the same sequence recognised by their REase partner. The recognition sequence is usually 4-8 bp in length and modifications consist typically of methylation of adenine or cytosine bases, taking place after replication of the prokaryotic genome. Thus, all genetic elements that do not contain the appropriate modifications are recognised as “foreign” by the REase and cleaved. Cleavage can occur either within

or at locations up to 1000bp from the recognition site, which requires ATP-dependent translocation of the RM complex on the DNA. Depending on the subunit combination, characteristics of the recognition and cleavage site and cofactor requirements, the RM systems can be classified into four main groups (I-IV, Tock and Dryden, 2005). Type IV systems do not contain an MTase activity, but instead recognise and cleave modified DNA substrates. Type III systems are often found in phage genomes.

In response to this system, a variety of escape mechanisms has been found in phages, viruses and conjugational plasmids. Anti-restriction strategies include, but are not limited to: i) encoding of or stimulation of the host MTase to methylate the phage genome in the same pattern as the host in order to avoid recognition; ii) loss or re-orientation of recognition sequences; iii) incorporation of unusual bases within their genome; iv) shielding of the recognition sequences upon injection by DNA-binding proteins encoded by the phage or plasmid; v) degradation of RE cofactors (e.g. S-adenosyl methionine or SAM); v) obstruction of RM enzymatic activity by encoding inhibitor proteins (e.g. DNA-mimicking proteins such as the Ocr protein of phage T7) (reviewed in Tock and Dryden, 2005). Additionally, RNA viruses evade this system. The fact that the RM systems are leaky or can be subverted can also be beneficial to the cell, in the sense that horizontal gene transfer can be a mechanism by which prokaryotes gain novel functional activities (Tock and Dryden, 2005).

1.1.2 Abortive infection

Abortive infection (Abi) is the common phenotype caused by a number of resistance mechanisms that can vary in their specific molecular mechanisms and have little or no evolutionary relationship. These mechanisms operate after the virus adsorption and injection of its genetic material into the cytoplasm, and can inhibit various stages of the virus life cycle inside the host cell, such as the transcription and replication of the virus/phage genome, protein production and virus assembly (Chopin *et al.* 2005). The common result is the complete inhibition of virus proliferation and death of the host cell. This “programmed suicide” is advantageous to the bacterial population as a whole since it prevents the spread of the infection, and for this reason it has been hypothesised that it represents a form of bacterial altruism (reviewed in Chopin *et al.* 2005; Hyman and Abedon, 2010). Abi systems are widespread among Gamma-proteobacteria, Actinobacteria, Cyanobacteria and Firmicutes, and many were isolated first in *lactococci* (Chopin *et al.* 2005). The majority are plasmid-encoded and frequently consist only of one gene. Abi mechanisms can be quite specific, with each mechanism targeting only certain groups of phages. Some of these mechanisms are under tight cellular regulation and their toxic activity is phage-induced (e.g. AbiD1, which inhibits the resolution of branched DNA structures), but others are constitutively expressed at low levels (e.g. AbiA, AbiB, AbiK) (Chopin *et al.* 2005).

It has been shown that a number of toxin-antitoxin (TA) systems also mediate abortive infection upon phage/virus infection among their other functions (e.g. inducing stasis or cell death as a stress response, ensuring maintenance of mobile genetic elements, regulating pathogenicity etc, reviewed in Van Melderen, 2010; Blower *et al.* 2011). TA cassettes consist of a single promoter controlling the expression of a gene pair, encoding for the unstable antitoxin and the stable toxin (figure 1.3). Both elements can be proteins or RNA molecules according to the system classification (Bukowski *et al.* 2011). Formation of a toxin-antitoxin complex inhibits toxin activity, which can manifest in a variety of mechanisms, but is either lethal or restrict cellular growth (Van Melderen, 2010; Blower *et al.* 2011). TA systems are widespread in bacteria as well as archaea (Makarova *et al.* 2009), although the three systems that mediate Abi have been characterised from bacteria.

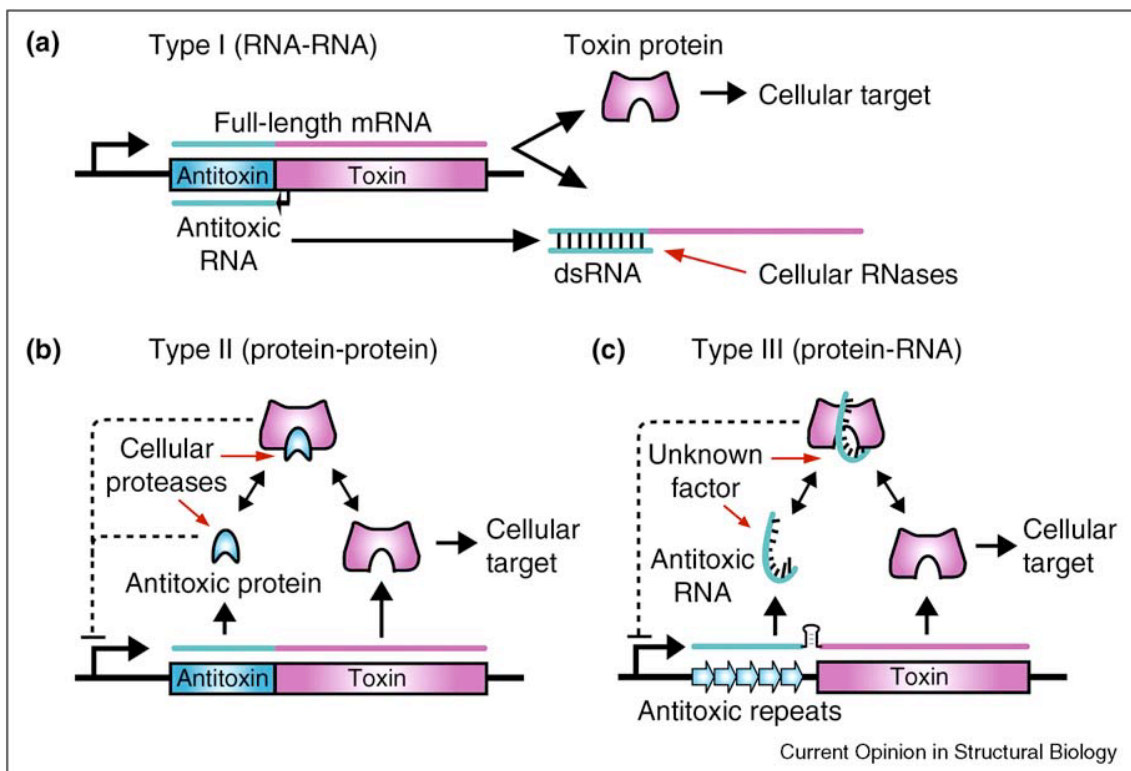


Figure 1.3: Outline of toxin-antitoxin systems

The two genes form a bicistronic operon. (a) In type I systems the toxin is a protein and the antitoxin is an antisense RNA. Silencing takes place at the RNA level. (b) In type II systems both the toxin and the antitoxin are proteins which can form a tight complex, preventing the action of the toxin. The antitoxin can get degraded by cellular proteases and release the toxin. (c) In type III systems the antitoxin is an RNA transcribed by an array of repeat sequences, which interacts and inhibits the protein toxin. Regulation is mediated by a transcriptional terminator (stem-loop structure) between the two genes (adapted from Blower *et al.* 2011).

In particular, it has been shown that TA cassette *mazEF* is stimulated by and can abort infection by phage P1 in *E. coli* by inhibiting translation, as toxin MazF is an mRNA interferase (Hazan and Engelberg-Kulka, 2004; Nariya and Inouye, 2008). Secondly, TA system *hok-sok* encoded in plasmid R1 of Gram-negative bacteria is stimulated by phage T4 and causes cell membrane damage by toxin Hok (Thisted and Gerdes, 1992). Finally, the plasmid-encoded ToxIN system has also been shown to mediate Abi in Gram-negative bacteria, but its mode of action is still unclear (Fineran *et al.* 2009).

Evidently, phages have developed mechanisms to overcome damage caused by Abi agents and avoid abortion of infection. This has been observed so far either by recombination with a cognate prophage or by point mutations within the phage genome (Chopin *et al.* 2005), however the nature of the Abi phenotype renders the characterisation of both the defence mechanisms and the anti-abortive phage strategies extremely difficult.

1.1.3 Dissemination, evolution and additional roles of defence systems

All the prokaryotic defence systems described here seem to share three common characteristics: they propagate mostly via horizontal gene transfer (HGT), exhibit extremely high rates of molecular evolution and display traits of selfish genetic elements (Stern and Sorek, 2010).

Evidence for the lateral mode of distribution is provided by the facts that a large number of these systems are encoded by plasmids, phages, prophages or hypervariable loci in the prokaryotic chromosome, and by homology found between distantly related strains. Often, they exhibit a codon usage bias and GC content different from the rest of the genome they reside in (Kobayashi, 2001; Gogarten *et al.* 2002; Chopin *et al.* 2005; Godde and Bickerton, 2006; Makarova *et al.* 2009). An obvious question is how HGT takes place and bypasses the pre-existing resistance mechanisms of a given organism. This is attainable due to the “leaky” nature of all the mechanisms discussed here, the fact that the rapid evolutionary rates enable invading elements to continuously develop anti-resistance strategies, and the fact that many of these systems are under tight control in certain development stages of the organism, or are subject to “phase variation” (Hoskisson and Smith, 2007). The latter refers to the alternate expression of different combinations of subunits in an RM system, or even the reversible inactivation of the system itself.

The extreme selection pressures acting on both host resistance and invader anti-resistance systems lead to rapid evolutionary rates, in what has been described as a co-evolutionary “arms race” (Hoskisson and Smith, 2007; Stern and Sorek, 2010). This is evident in the wide array of classes and types of RM mechanisms, systems

mediating Abi and CRISPR/Cas subtypes (discussed later), and even hypervariability of the target recognition domains in RM systems.

Inherent characteristics of certain genetic elements can lead to an increase of their relative frequency within a given population. These types of genetic elements are characterised as “selfish”. This theory has been put forward to explain the behavior of several prokaryotic defence mechanisms, especially the RM systems (Kobayashi, 2001) and TA modules (Makarova *et al.* 2009). The following observations support this theory: i) loss of these systems results has deleterious effects on the host cell (figure 1.4 A); ii) if a competing genetic element enters the cell (e.g. another RM or TA system), competitive exclusion between the two will lead either to attack and destruction of the invader or to host death, thus preventing the competing system from establishing itself in the population (figure 1.4 B); iii) these systems display extreme mobility between genomes.

As a consequence, it can be argued that participation of these systems in cellular defence is primarily a self-maintenance strategy, that happens to be advantageous to the host cell (Kobayashi, 2001; Makarova *et al.* 2009; Stern and Sorek, 2010).

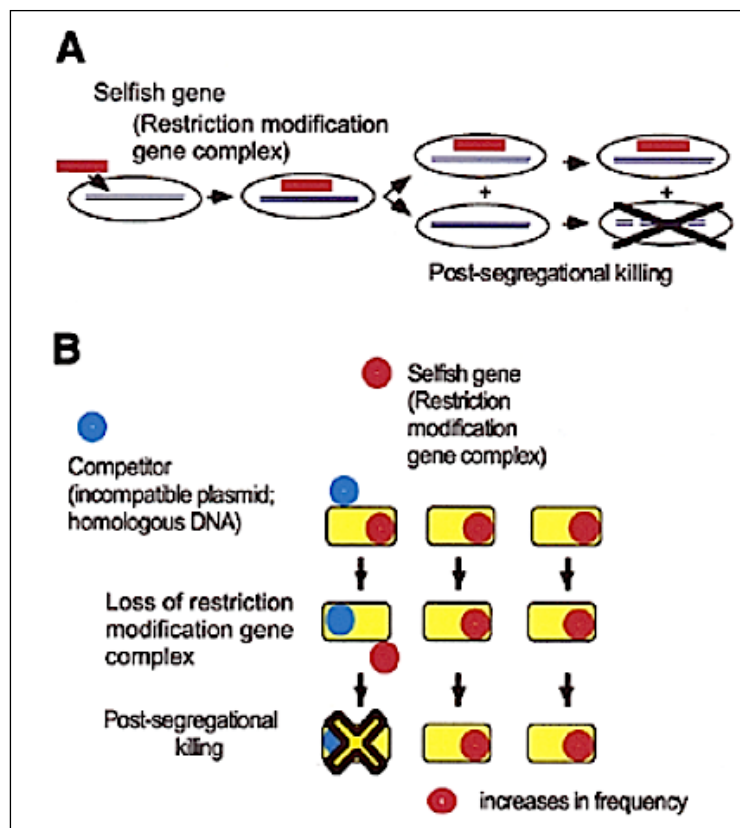


Figure 1.4: Characteristic behavior of selfish genetic elements

(A) Post-segregational killing of the carrier that loses the element, establishing its maintenance in the population. (B) Competitive exclusion between two equally deleterious selfish elements (adapted from Kobayashi, 2001).

All of the prokaryotic resistance mechanisms described here seem to be implicated in additional pathways within the host cell. This has been extensively studied in TA systems, which are known to be activated in response to various types of cellular stress and also maintain genomic integrity by preventing loss of certain mobile genetic elements (Van Melder, 2010). The induction modes and reversible effects of certain Abi systems has also led to the suggestion that they might also serve other cellular functions (Chopin *et al.* 2005). In several RM systems, following loss of the REase activity the MTase adopts a regulatory role in various aspects of DNA metabolism. Moreover, the “phase variation” controlled expression of RM systems is linked to several regulatory functions, such as allowing for differential epigenetic modifications of the genome, enhancing pathogenicity and enabling acquisition of foreign DNA (reviewed in Stern and Sorek, 2010). Putative alternate roles of the CRISPR system will be discussed later in this chapter. This versatility of prokaryotic defence systems seem to be an important and defining characteristic, highlighting the extreme plasticity and dynamic nature of the prokaryotic genome and its exceptional ability to adapt.

1.2 Discovery and characterisation of the CRISPR arrays in prokaryotic genomes

In contrast to eukaryotic genomes, less than 5% of the genome of most prokaryotic phyla comprises of repetitive DNA (Ussery *et al.* 2004). Therefore, it has been proposed that its existence offers certain advantageous characteristics to the carrier, since it survived natural selection over evolutionary time.

A distinct family of direct repetitive DNA sequences in prokaryotic genomes are the clustered regularly interspaced short palindromic repeats (CRISPR). This family was first identified as a distinct class of interspersed short sequence repeats (SSR), adjacent to the isozyme-converting alkaline phosphatase (*iap*) gene by Ishino *et al.* in 1987 and Nakata *et al.* in 1989 in *Escherichia coli* K12. The same class of repeats was found soon in other prokaryotic species such as *Mycobacterium tuberculosis* (Hermans *et al.* 1991), *Haloferax mediterranei* (Mojica *et al.* 1995), and *Thermotoga maritima* (Nelson *et al.* 1999). They were recognized as a defined prokaryotic family of short regularly spaced repeats (SRSR) by Mojica *et al.* in 2000, who in agreement with Jansen *et al.* introduced the acronym CRISPR in an initial study of the CRISPR-associated system in 2002.

The main, conserved features of the CRISPR system are the following, outlined in figure 1.5 (reviewed in Sorek *et al.*, 2008; van der Oost *et al.* 2009; Horvath and Barrangou, 2010; Karginov and Hannon, 2010; Marraffini and Sontheimer, 2010; Deveau *et al.* 2010; Al-Attar *et al.* 2011):

- i. These elements consist of direct repeat sequences which range in size from 21-48 bp (with an average size of 32bp) and in number from 2-375 repeats per locus (with an average of 27). The repeat sequences can be partially palindromic, in the form of inner and terminal imperfect inverted repeats of up to 11bp (Godde *et al.* 2006). CRISPR loci are usually homogenous in their repeat sequence. In terms of sequence conservation, phylogenetically distant species generally show greater variation of the repeat sequences than closely related species, although many exceptions have been observed (Jansen *et al.* 2002). The repeats can be divided into 12 clusters based on sequence similarity and secondary structure formation (Kunin *et al.* 2007). Six of these clusters exhibit high and intermediate RNA folding scores indicating that the repeats, when transcribed, form stable secondary stem-loop structures mediated by the palindromic sequences, hypothetically facilitating recognition by CRISPR-associated proteins. Moreover, some of the clusters contain the conserved sequence GAAA(C/G) in their 3'-terminus, indicating a possible protein binding site. .
- ii. The repeat sequences are regularly spaced by unique intervening sequences (spacers) of variable length, which range in size from 26-72 bp. A fraction of the spacer sequences was found to exhibit significant similarity to sequences from phage DNA and conjugative plasmids, with the highest degree of similarity for a given spacer found within genetic elements associated to the carrier. Taking into account the limited number of characterized viral genomes and conjugative plasmid sequences, it was concluded that the spacer sequences originate from these foreign genetic elements (Mojica *et al.* 2005, Pourcel *et al.* 2005). In support of this theory, 40% of the spacers in lactic acid bacteria CRISPR loci were found to be homologous to streptococci phage genomes and the respective conjugative plasmids (Bolotin *et al.* 2005). Crenarchaeal CRISPR spacers yield matches to fuselloviruses, rudiviruses and β -lipothrixviruses. Spacer sequence matches were found in both the sense and anti-sense strands and both gene coding and intergenic regions of phage genomes (Shah *et al.* 2009). The viral or plasmid sequence that is complementary to a given spacer sequence is known as a "protospacer"
- iii. Leader sequences of a size order of 100-550 bp have been detected in association with several (but not all) CRISPR loci. They are located directly upstream of the cluster, with respect to the strand orientation of the repeat sequence. These sequences appear to have a high A-T content, are rich in homopolynucleotide regions, lack open reading frames and are generally not conserved between distantly related species (Jansen *et al.* 2002), but exhibit similarity between related species. Analysis of the primary transcripts of CRISPR loci in several species revealed that transcription initiates within the leader region (Hale *et al.* 2009;

Lillestol *et al.* 2006), and putative promoter motifs were identified in leader regions of *Sulfolobus acidocaldarius* (Lillestol *et al.* 2006) and *E. coli* K12 (Pul *et al.* 2010) confirming that these regions act as transcription promoters for the sense strand of the CRISPR arrays. Moreover, it was initially deduced by comparative analysis (Lillestol *et al.* 2006) and subsequently confirmed by genetic studies in *Streptococcus thermophilus* (Barrangou *et al.* 2007) that novel spacers are incorporated along with a novel repeat into the leader proximal end of the CRISPR loci. Therefore, leader regions seem to be playing the dual role of controlling CRISPR transcription and the growth of the array, by interacting with the appropriate proteins for the addition of new spacers.

iv. A number of protein families have been designated CRISPR-associated (Cas), and together with the repeat cluster are regarded as a unified system (Jansen *et al.* 2002; Haft *et al.* 2005; Makarova *et al.* 2006). These families are present only in CRISPR containing species, located adjacent to the repeat cluster with a generally conserved orientation. No homologues of the *cas* genes were found in eukaryotic or CRISPR-negative genomes. Only one set of *cas* genes is present in species carrying multiple CRISPR loci with the same repeat sequence, but if multiple loci with varied repeat sequence are present, then a respective number of *cas* gene sets is observed. A *cas* gene region can comprise of as many as 20 different, tandem-arranged genes with no preferential direction of their reading frames, and can be found on either side of a CRISPR locus. An analysis of the Cas genes will be presented in the following section.

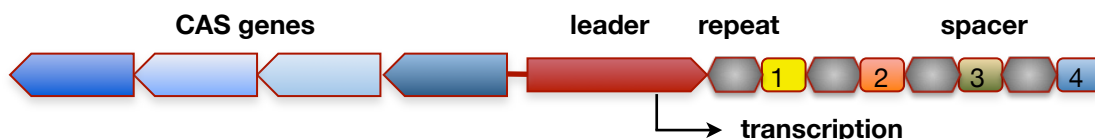


Figure 1.5: Graphic representation of a CRISPR locus and the adjacent *cas* gene operon in a prokaryotic genome

Cas genes are depicted as blue arrows, the leader sequence in red, repeats as dark grey boxes and interspersing spacers as colored, numbered boxes. Direction of transcription is indicated by the black arrow.

The number of CRISPR loci per genome ranges from 0 to 20, with *Methanocaldococcus jannaschii* containing the highest number found to date (Jansen *et al.* 2002, Godde *et al.* 2006, Lillestol *et al.* 2006). According to most recent reviews CRISPR loci are present in 90% of the sequenced archaeal genomes, covering both phyla of Crenarchaeota and Euryarchaeota, and in 40% of the sequenced bacterial genomes, adding up to 639 genomes analysed up to date (table 1.1, Makarova *et al.*

2011a). It has been observed that archaeal clusters, especially from thermophilic organisms, are in general multiple and larger than the bacterial ones, and can represent up to 1% of the prokaryotic genome. Clusters are also present in archaeal conjugative plasmids, such as pNOB8 and pKEF9 of *Sulfolobus sp* and bacterial megaplasmids such as pTT27 of *Thermus thermophilus* (Lillestol *et al.* 2006, Godde *et al.* 2006).

Taxonomic group	Genomes analysed	Genomes containing cas1	proportion of cas1-containing genomes (%)	type I system	type II system	type III system
Archaea						
Crenarchaeota	17	15	88	15	0	16
Euryarchaeota	47	37	79	33	0	23
Total	67	54	81	50	0	40
Bacteria						
Actinobacteria	72	26	36	28	15	8
Aquificae	7	5	71	7	1	4
Bacteroidetes-Chlorobi group	32	16	50	14	2	6
Chlamydiae–Verrucomicrobia group	10	2	20	0	1	1
Chloroflexi	10	9	90	9	2	7
Cyanobacteria	14	7	50	7	1	7
Firmicutes	126	56	44	40	17	23
Proteobacteria	318	107	34	117	20	22
Spirochaetes	13	3	23	2	1	0
Thermotogae	11	10	91	10	0	9
Total	639	256	40	245	65	99

Table 1.1: Taxonomic distribution of CRISPR-Cas systems in 706 analysed genomes

Different CRISPR system types can coexist in different genomes. Adapted from Makarova *et al.* 2011a.

The origin of spacer sequences and the analogies observed by Makarova *et al.* (2006) between the system components and the eukaryotic RNA interference led a number of groups to propose that the CRISPR loci and their associated genes represent an adaptive prokaryotic resistance system against infections from extrachromosomal elements, functioning via RNA interference (Mojica *et al.* 2005, Pourcel *et al.* 2005, Bolotin *et al.* 2005, Makarova *et al.* 2006). Moreover, Mojica *et al.*

(2005) had already combined reports of foreign genetic elements such as viruses and conjugative plasmids failing to infect CRISPR-carrier strains whose spacers exhibited homology with these elements. The first experimental validation of the CRISPR function was achieved in 2007 by Barrangou *et al.* when CRISPR loci of *Streptococcus thermophilus* were shown to incorporate new spacers homologous to phage genomic sequences during the generation of phage-resistant mutants, and resistance specificity was shown to depend on the spacer sequence content. In the same study, knockout of two cas genes resulted in loss of phage resistance and inability to incorporate new spacers respectively, confirming the functional association of the cas genes with the CRISPR elements (Barrangou *et al.* 2007).

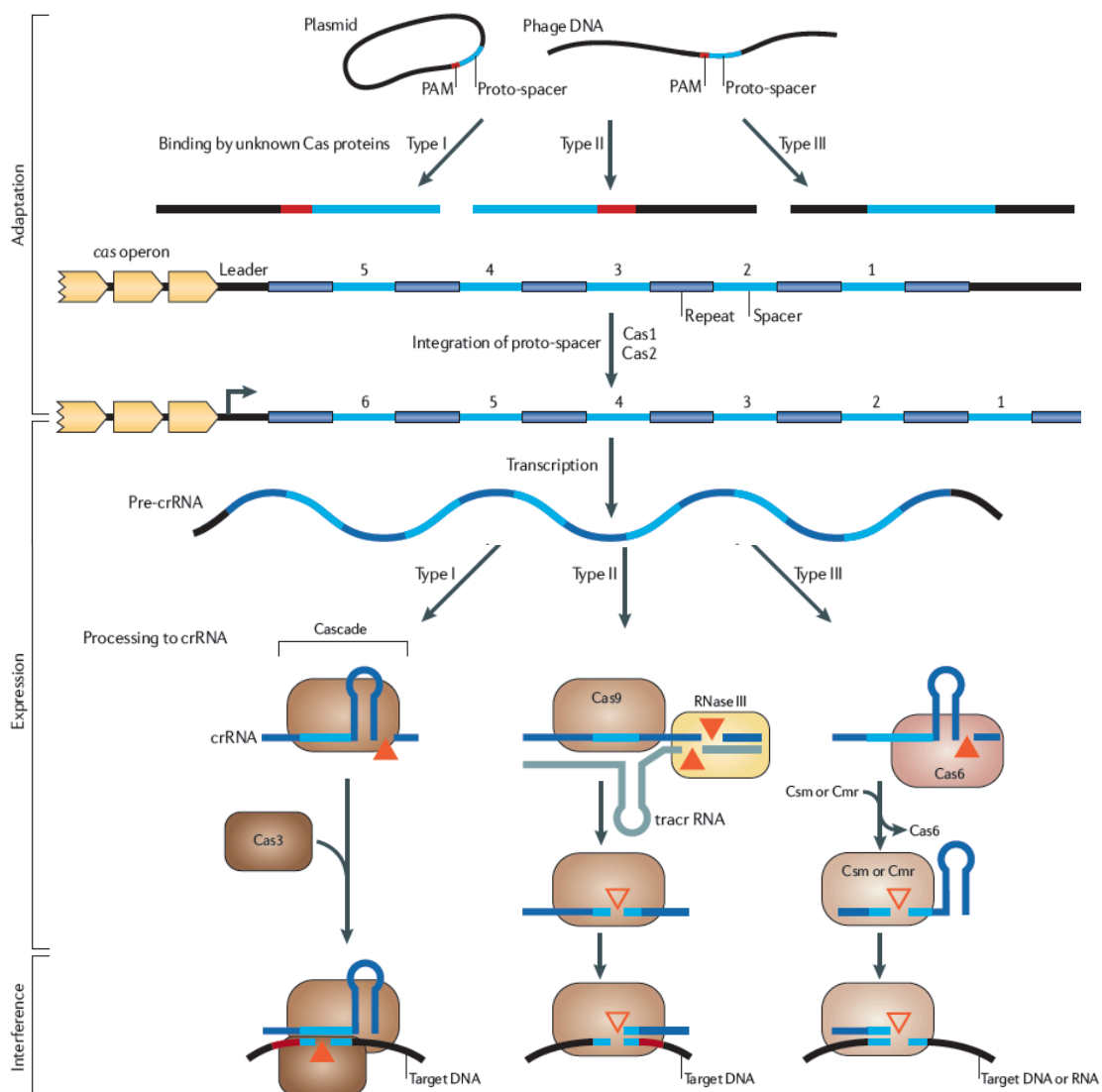


Figure 1.6: Outline of the CRISPR/Cas mode of action

The current model for the three stages (adaptation, expression and interference) of CRISPR functioning for each subtype, as inferred by genetic and biochemical studies discussed in this chapter (adapted from Makarova *et al.* 2011a).

According to the proposed mechanism (Makarova *et al.* 2006), the entire CRISPR repeat region is theoretically transcribed as a single primary transcript, and after a series of processing steps small interfering antisense RNA molecules of the size of a repeat/spacer unit (referred to as mature crRNAs or psiRNAs in the literature) are produced. This procedure could be under regulation by Cas proteins and induced by stress or phage infection. The mature psiRNA molecules could then anneal to the respective foreign mRNA, resulting in translation repression or cleavage of the dsRNA molecule, thus silencing the foreign genes and inhibiting phage or plasmid proliferation (figure 1.6). The Cas proteins are proposed to comprise the protein machinery of this immune system, mediating the generation and maintenance of the CRISPR loci, the processing and integration of new spacers as well as the RNA silencing process. The discrete stages of this mechanism will be discussed in more detail in subsequent sections.

1.3 CRISPR-associated protein families and current classification of the CRISPR/Cas system

The neighborhood of *cas* genes (comprising of more than 20 genes) was initially identified and characterized by Makarova *et al.* in 2002 by genomic context analysis, but it was wrongly predicted to be a novel DNA repair system specific for thermophiles, as no connection with CRISPR was detected at the time. Almost simultaneously, Jansen *et al.* identified by *in silico* analysis four genes located in the vicinity of CRISPR loci that were designated CRISPR-associated (*cas*1-4; Jansen *et al.* 2002). The first protein found to bind to CRISPR loci was a genus-specific uncharacterized protein in *Sulfolobus* species corresponding to *sso454* (Peng *et al.* 2003), recognizing double and single repeat DNA sequences and producing an opening on the opposite side. Haft *et al.* in 2005 identified a guild of 45 Cas protein families by Hidden Markov models, a categorization refined by Makarova *et al.* in 2006 taking into account genomic context information, resulting in 25 Cas protein families (Makarova *et al.* 2006). These families are proposed to be involved in the generation, expansion, maintenance, transfer between genomes and function of the CRISPR elements.

With the rapid growth of experimental characterisation and identification of novel CRISPR systems in more prokaryotic genomes, it became apparent that existing CRISPR/Cas classification systems grew increasingly inadequate and did not reflect the emerging phylogenetic relationships between the system components. Moreover, with the elucidation of many Cas protein structures from different families and analysis of an increasing number of gene sequences, previously undetected homologous relationships emerged which enabled the unification of certain Cas families and the

identification of novel ones (Makarova *et al.* 2011b). As a result, recently Makarova and colleagues (2011a) proposed an updated, polythetic classification of CRISPR/Cas systems based on gene composition, operon organisation and the phylogenetic and functional relationships between Cas genes. According to the novel classification, CRISPR/Cas systems are organised into three phylogenetically distinct types (I-III), and each major type can be further divided into individual subtypes (Makarova *et al.* 2011a and b). This classification is summarised in figure 1.7 and the subtypes distribution in table 1.1.

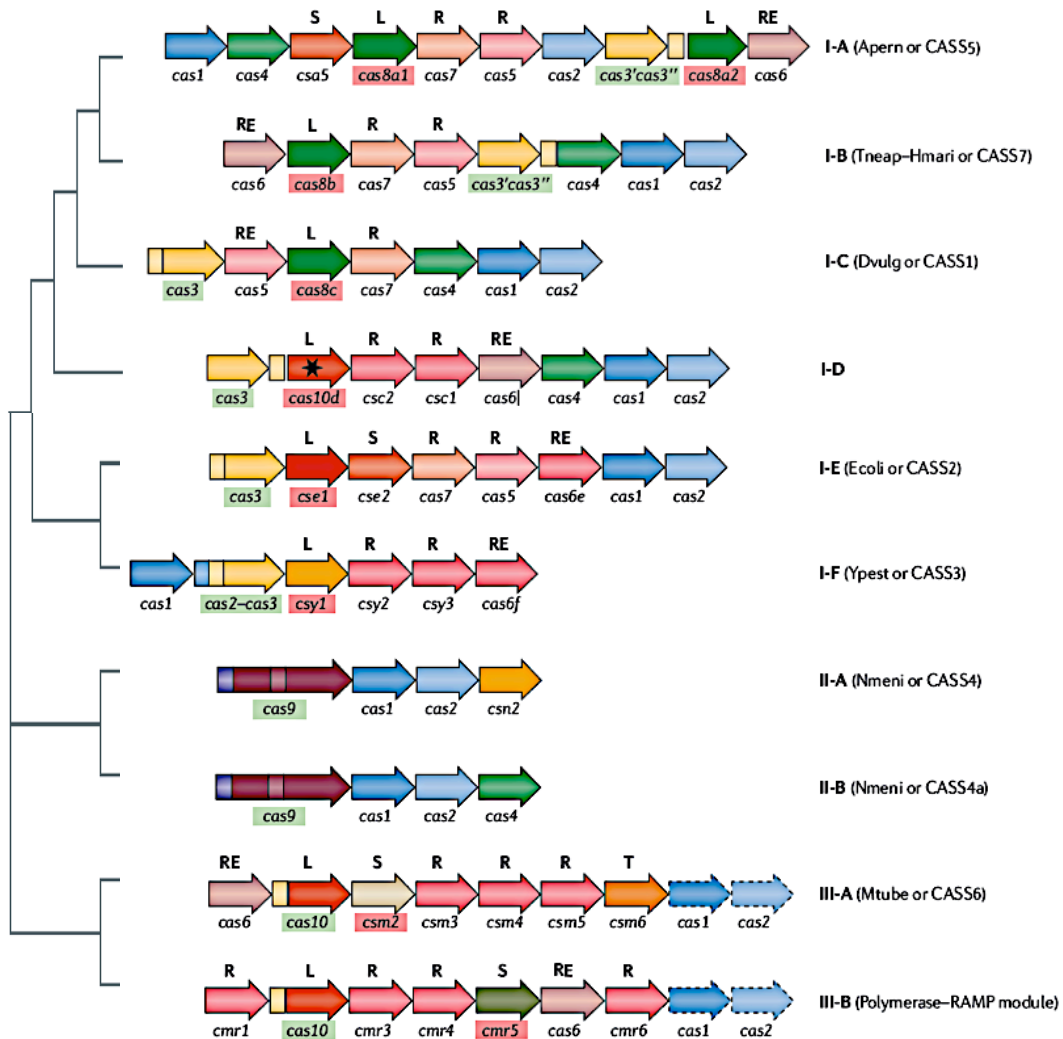


Figure 1.7: Outline of the main types and subtypes of the CRISPR/Cas systems and their phylogenetic relations

The most common composition and arrangement of cas genes is shown for each subtype, but gene order may vary in each organism. Gene families are color-coded and the family name can be seen under each gene. Signature genes for each main type are highlighted in green, and for each subtype in red. The star in gene *cas10d* indicates a putative inactivated polymerase - HD domain. The letters above certain genes stand for: RE: processing endonuclease for crRNA maturation; L: large subunits of effector complexes mediating interference; S: small subunits of effector complexes; R: subunits of effector complexes that belong to the RAMP superfamily (Repeat Associated Mysterious Proteins; described in chapter 3). Dashed genes in type III systems may not be part of the same operon. Adapted from Makarova *et al.* 2011a.

The three main CRISPR/Cas types share a common core of two genes, *cas1* and *cas2*, which are highly conserved and are found in almost all CRISPR-containing species. Cas1 a highly conserved, basic protein that belongs to COG1518 (all COG groups mentioned in this text refer to the analysis performed by Makarova *et al.* 2002). Comparative sequence analysis and certain conserved residue patterns indicate that it might be a putative novel nuclease and/or integrase (Makarova *et al.* 2002). Metal-dependent nuclease activity on ss/ds DNA (non-sequence specific) was confirmed by Wiedenheft *et al.* (2009) along with the elucidation of the Cas1 structure from *P. aeruginosa* which revealed a unique fold (figure 1.8). Additionally, Cas1 from *S. solfataricus* exhibited a high binding affinity for ss/ds DNA, ss/ds RNA and DNA-RNA hybrids, as well as strand annealing activity (Han *et al.* 2009).

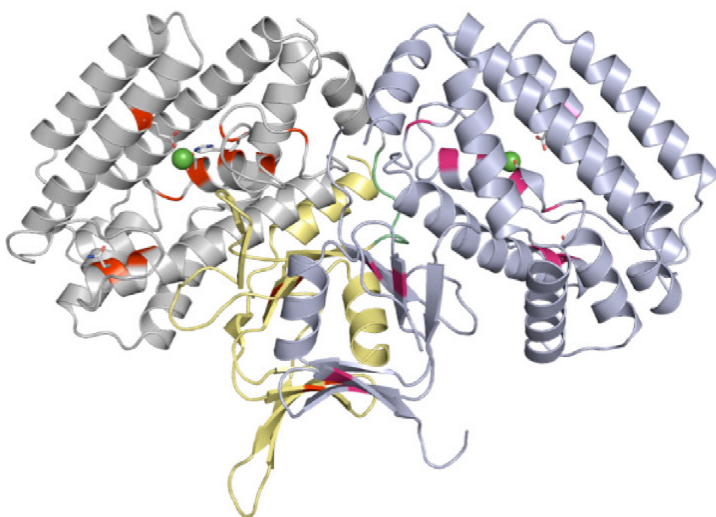


Figure 1.8: Crystal structure of Cas1

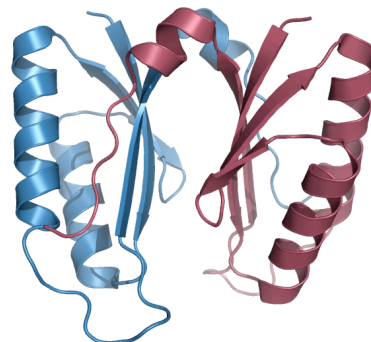
Cartoon representation of the *P. aeruginosa* Cas1 homodimer (adapted from Wiedenheft *et al.* 2009). The N-terminal domain of chain A is colored in yellow, and the C-terminal α -helical domain which contains the active site in gray. Chain B is colored in light blue. Conserved residues making up the active site are in red. Three of the residues (E190, H254 and D268) coordinate a manganese ion (green sphere).

The *cas2* gene encodes a small (80-120aa) protein member of COG1343. Distant similarities were found between members of this COG and a class of sequence-dependent, single-strand RNA nucleases called PIN-domain nucleases (after their identification in the N-terminus of the pilin biogenesis PilT protein), leading to the speculation that Cas2 might also possess ribonuclease activity (Makarova *et al.* 2006). The structure of Cas2 from *S. solfataricus* was solved by Beloglazova *et al.* (2008) revealing an RRM-like domain (RNA recognition motif; structural motif consisting of four β -strands and two helices arranged in a α/β sandwich) (figure 1.9), while the protein exhibited metal-dependent ssRNAse activity. The universal distribution of this gene pair along with experimental evidence discussed in subsequent paragraphs, has led to the assumption that Cas1 and Cas2 mediate the integration of novel spacer sequences into the CRISPR loci (reviewed in Sorek *et al.* 2008; van der Oost *et al.* 2009; Horvath and Barrangou, 2010; Karginov and Hannon, 2010; Marraffini and

Sontheimer, 2010; Deveau *et al.* 2010; Al-Attar *et al.* 2011). The role of these core proteins in the current scheme of the CRISPR mode of action will be discussed later.

Figure 1.9: Crystal structure of Cas2

Structure of Cas2 from *S. solfataricus*, solved by the SSPF (PDB code: 2IVY). The active conformation is a homodimer, with the interface formed by the tandem β -sheets in each monomer that make up the RRM motif. Conserved residues are located on the loops at the edge of the central cleft, at the bottom of the structure.



Type I systems are characterised by the presence of *cas3* (COG1203), a gene encoding for a protein with conserved superfamily II helicase motifs and an additional HD-nuclease domain, encoded separately in certain subtypes (Makarova *et al.* 2002). Type I systems also contain multiple representatives of the RAMP superfamily (Repeat associated mysterious proteins), which are suggested to form large heteromeric complexes and take part in invader silencing (Brouns *et al.* 2008). The RAMP superfamily encompasses a large variety of protein families with ferredoxin-like folds, predicted to have RNA-binding activity (Makarova *et al.* 2002, 2006; Haft *et al.* 2005) and will be discussed in more detail in chapter 3. Characteristic RAMP families associated with type I subtypes include Cas5, Cas6 and Cas7 (COG1857) protein families (Makarova *et al.* 2011a). Cas6 has been shown to possess metal-independent, sequence specific RNase activity, and is the processing endonuclease that generates the mature interfering RNA units (referred to as crRNAs from now on) from the primary CRISPR transcript, in every type/subtype it is associated with. An additional protein found in four out of six type I subtypes and a type II subtype is Cas4 (COG1468), a member of the RecB exonuclease family (Jansen *et al.* 2002, Makarova *et al.* 2002). A number of studies have concluded that the targets of type I systems are DNA viruses and plasmids (among others Brouns *et al.* 2008; Marraffini *et al.* 2008, Garneau *et al.* 2010).

Type II systems have been found only in bacteria and contain only the signature gene *cas9* (COG3513), the core *cas1/cas2* genes and either *cas4* or *csn2*, a modular gene (Makarova *et al.* 2011a). Cas9 family members are predicted to be large (about 1000 residues), multidomain proteins including an N-terminal RuvC-like domain (RuvC is a Holliday junction resolvase that belongs to the RNase H fold; Aravind *et al.* 2000) and an HNH nuclease domain, common in restriction endonucleases (Makarova *et al.* 2002). Targeting of plasmid and phage DNA was demonstrated *in vivo* for this

system, and Cas9 is implicated in the interference stage although no biochemical characterisation has been presented (Barrangou *et al.* 2007; Garneau *et al.* 2010).

Type III systems are characterised by the presence of *cas10* (COG1353). Among the identified domains of this large multidomain protein (~1000 residues) is a permuted HD-superfamily hydrolase near the N-terminus, a globular uncharacterised $\alpha+\beta$ domain, a Zinc-ribbon (well-known nucleic acid interacting domain) and the core palm domain of DNA/RNA polymerases and nucleotide cyclases near the C-terminus (Makarova *et al.* 2002, 2006). The function of this protein is yet to be elucidated, but it has been shown to form multimeric complexes with the additional RAMP Cas proteins in type III-B operons which can effectively target RNA *in vitro* (Hale *et al.* 2009). Targeting of DNA has also been demonstrated *in vivo* for type III-A systems (Marraffini and Sontheimer, 2008). *cas6* is also part of type III systems. The core *cas1* and *cas2* genes are occasionally missing from type III operons, but in these cases they are found to co-exist with other CRISPR/Cas systems (type I or type II) encoding *cas1* and *cas2* in the same genome. This supports the theory that *cas1* and *cas2* are involved in a different stage of CRISPR functioning, and co-regulation is not necessary (Makarova *et al.* 2011a). Mechanistic details of each stage in every CRISPR/Cas type will be discussed in detail in subsequent sections.

1.4 The three stages of the CRISPR/Cas mode of action

The CRISPR/Cas system functioning can be divided into three mechanistically distinct stages (Makarova *et al.* 2006; van der Oost *et al.* 2009). The first stage involves the first encounter with the invader extrachromosomal element, the selection of the protospacer among the invader DNA sequences and the incorporation of the invader-derived short DNA sequence into the CRISPR array as a novel spacer. The second stage, termed CRISPR expression, consists of the transcription and processing of the CRISPR arrays to generate mature crRNAs (CRISPR RNAs), which are bound by Cas effector proteins and serve as guide sequences for the third stage of CRISPR-mediated defence. During this final stage, the crRNA-Cas protein complexes recognize, bind and inactivate the invading virus or plasmid, most likely by direct Cas-mediated degradation of the target nucleic acid. The current state of understanding of the molecular mechanisms and protein machinery taking part in these three stages are described below, and summarised in figure 1.6.

1.4.1 Stage I: Spacer selection and integration into CRISPR arrays

This stage is also known as the adaptation stage of CRISPR functioning, as the incorporation of novel spacers enables the swift adaptation of prokaryotic cells to the

dynamic environmental pool of mobile invader genetic elements. The growing CRISPR arrays serve as heritable “libraries” of past infective events that render their carrier immune to subsequent attacks by previously encountered viruses or plasmids.

The first experimental evidence of novel spacer acquisition and the implications for CRISPR mediated anti-viral defence was provided by Barrangou *et al.* (2007) in *Streptococcus thermophilus* (type II CRISPR/Cas system). In this study it was demonstrated that during the natural generation of resistant mutants by phage challenging, the surviving mutants contained novel phage-specific spacers in their CRISPR loci. Thus it was proven that the cells adapt to the new threat by altering their CRISPR loci to accommodate invader-derived sequences. The new spacers originated from both strands of the phage genome, without preferential targeting of coding or intergenic regions, providing also the first indication that DNA is the ultimate target. The level of resistance against a single pathogen correlated with the number of spacers acquired from that particular pathogen, and also with the level of identity between the protospacer/spacer sequences, as only the spacers without mismatches conveyed resistance. Deletion of the novel spacers resulted in increased sensitivity of the produced strains, thus establishing the link between CRISPR spacer content and phage resistance. This was also the first study to demonstrate the direct involvement of the Cas genes in the integration process, because inactivation of *csn2* rendered the mutant unable to acquire new spacers.

Since then, apart from additional studies of the *S. thermophilus* system (Deveau *et al.* 2008, Horvath *et al.* 2008), adaptation by spacer incorporation has either been shown to occur naturally in a number of species such as *Streptococcus mutans* (van der Ploeg *et al.* 2009), or inferred by comparative genomic analysis of closely related strains (among others: *Yersinia pestis*, Pourcel *et al.* 2005; *Thermotoga* sp. DeBoy *et al.* 2006; *Thermococcales*, Portillo *et al.* 2009; *Sulfolobales*, Lillestol *et al.* 2009, Held *et al.* 2010; lactic acid bacteria: *Lactobacillales* and *Actinobacteria*, Horvath *et al.* 2009), and metagenomic analysis of natural microbial populations (Tyson *et al.* 2008; Andersson *et al.* 2008; Heidelberg *et al.* 2009). A common observation in all studies was the polarized addition of new spacers at the leader-proximal end of the CRISPR locus along with the duplication of a repeat sequence, leading to the expansion of the array by a complete repeat-spacer unit (Pourcel *et al.* 2005; Lillestol *et al.* 2006; Andersson and Banfield, 2008; Tyson and Banfield, 2008). A consequence of this is that comparison of the spacer content of CRISPR arrays reflects shared ancestry between related strains and can be used to reconstruct their recent evolutionary history and monitor virus/host interaction dynamics in microbial communities. The growth of CRISPR loci is also controlled by deletion of spacer regions via internal recombination events, as the unlimited expansion of the arrays is unsustainable for the cells. Internal recombination results in preferential deletion of

older spacers, which correspond to past infection events and are presumably less important for the survival of the organism (Horvath *et al.* 2008, Tyson and Banfield 2008).

The fact that the incorporation of new spacers occurs at the leader-proximal end of the loci suggests a role for the leader sequence other than transcription promoter (Lillestol *et al.* 2006; Marraffini and Sontheimer, 2008). It is possible that it contains binding sites for Cas/host proteins involved in spacer integration and repeat duplication. Interestingly, comparative analysis of the leader sequences in *Sulfolobales* revealed a series of more or less conserved motifs, albeit of low sequence complexity, with a different content/arrangement for each CRISPR family (Lillestol *et al.* 2009), of yet unknown significance.

The molecular mechanism of this stage is currently the least understood part of CRISPR biology. Due to the architecture of the CRISPR arrays it has been proposed that the insertion of new spacers proceeds through homologous recombination with the genomic CRISPR region, accompanied by (or preceded by) a repeat duplication event (Makarova *et al.* 2006). Initially it was also suggested that the putative reverse transcriptase function of the gene now classified as *cas10* could play a role in acquiring spacers from RNA sources, but with the elucidation of its role in the Cmr complex (multiprotein complex mediating interference in type III systems, described later) this proposition was discarded.

One of the key proteins thought to be involved in this stage is core protein Cas1, due to the following reasons: i) the universal distribution among CRISPR/Cas subtypes, indicating its essential role for the system, ii) initial bioinformatics analysis classifying Cas1 as a putative “nuclease/integrase” (Makarova *et al.* 2006), iii) the fact that in all subtypes studied until now it has not been found to associate with the protein machinery performing the crRNA processing and target interference (Brouns *et al.* 2008; Hale *et al.* 2009), and iv) its deletion has no effect on the second and third stage of CRISPR interference (Brouns *et al.* 2008). Experimental data on Cas1 homologs have so far been somewhat contradictory, but overall seem to support this hypothesis. The first biochemical and structural studies of Cas1 were performed by Wiedenheft *et al.* and Han *et al.* (2009) with orthologs from *Pseudomonas aeruginosa* and *Sulfolobus solfataricus* P2 respectively. Cas1 from *P. aeruginosa* was shown to be a metal-dependent sequence unspecific DNA endonuclease, capable of recognizing and cleaving ss and ds DNA substrates independent of their methylation pattern (Wiedenheft *et al.* 2009). The generated ds DNA products had an average size of 80bp which is much longer than the average spacer length for this organism (32 bp), suggesting that Cas1 interacts with additional Cas/host proteins in order to complete this step. In contrast, no nuclease activity was detected for the SsoCas1 ortholog, which appeared to bind ss/ds DNA, ss/ds RNA and DNA-RNA hybrids with

comparable affinities in the nanomolar range. Moreover, the protein displayed strand annealing activity in the presence of magnesium ions (Han *et al.* 2009), indicating its potential role in the final stages of spacer integration, when recombination-like events are likely to take place. The functional state of both Cas1 orthologs in solution was shown to be a dimer (figure 1.8 A). The latest study on Cas1 from *E. coli* (Babu *et al.* 2011) confirms that it is a multifunctional metal-dependent, nuclease which, apart from linear ss/ds DNA and ssRNA substrates, can also cleave branched DNA oligonucleotides such as Holliday junctions, replication forks and 5'/3' flaps in a sequence independent manner. These types of branched and cruciform-like substrates normally represent intermediates of DNA repair and recombination, and could potentially be formed by the palindromic repeat sequences within CRISPR repeats. Nuclease assays showed that the protein generated multiple cleavage products ranging in size from 5 nt to 34 nt and had a lower substrate size requirement than *P. aeruginosa* Cas1 as it was unable to cleave 60 bp DNA duplexes. The ability of *E. coli* Cas1 to cleave short branched substrates might be essential in the final steps of spacer integration, if indeed this step proceeds via recombination (Mojica *et al.* 2009). Interestingly, it was demonstrated that in *E. coli* Cas1 interacts physically with CasC (Cse4) and CasE (Cse3), two subunits of a multiprotein complex involved in the last stage of target interference termed CASCADE (CRISPR-associated complex for antiviral defence, described later), suggesting a previously unconsidered CASCADE involvement in the adaptation stage, and physically and genetically with the DNA repair/recombination associated proteins RecB, RecC, RuvB and UvrC, triggering hypotheses about the possibility that certain Cas components participate in DNA repair pathways (Babu *et al.* 2011). Overall, the current state of research on Cas1 seems to support its key role in the adaptation stage of CRISPR functioning, as well as pointing out the need to identify its functional partners.

Fusion or conserved gene synteny between Cas1 and Cas4 has been observed in a large number of genomes harboring type I systems, indicating a potential functional as well as physical association of these core system proteins (Makarova *et al.* 2006). Cas4 is one of the first CRISPR proteins for which a function prediction could be made, as it features conserved motifs characteristic of the RecB family exonucleases including a cysteine-rich motif responsible for DNA binding. The *E. coli* RecB protein is associated with recombinational DNA repair as a subunit of the larger RecBCD recombinase complex, and this role could be consonant with a theoretical recombination mechanism for spacer integration (Al-Attar *et al.* 2011). In type II-A systems it is proposed that *csn2* is the functional analogue of *cas4*, and indeed inactivation of this gene in *S. thermophilus* resulted in inability to acquire new spacers. (Barrangou *et al.* 2007).

The final protein proposed to take part in this stage is Cas2, which together with Cas1 form the core of the three distinct types of CRISPR systems. Deletion of the *cas1-cas2* gene pair did not have any effect on target interference in *E. coli*, and Cas2 has not been shown to interact strongly with any other Cas protein (Beloglazova *et al.* 2008). Biochemical characterisation of Cas2 homologs from a number of species including *S. solfataricus* P2 revealed that they are metal-dependent endoribonucleases specific for ssRNA substrates, cleaving preferentially in U-rich regions (Beloglazova *et al.* 2008). In agreement with the biochemical data, elucidation of the crystal structure of SsoCas2 (Sso1404) suggested that these proteins contain a $\beta\alpha\beta\beta\alpha\beta$ RNA-recognition motif (RRM) and their physiological state is a homodimer with a central cleft formed by the tandem arrangement of the β -sheets of each monomer (figure 1.9). Cas2 proteins were classified as a novel superfamily of the ferredoxin-like fold, as they are the first characterised nucleases to adopt this fold. The fact that neither Cas1 nor Cas2 exhibit any type of sequence or structure specificity in terms of substrate preference remains incomprehensible, as some mechanism of protospacer selection must exist, either in the form of a protospacer adjacent motif (PAMs: conserved di- or tri-nucleotide motifs associated with protospacers; described in 1.4.4) or a different sequence characteristic. This also renders necessary the existence of additional Cas components taking part in the adaptation stage, potentially by interacting specifically with the nucleic acids involved, recruiting Cas1/Cas2 or even fine-tuning their function by allosteric regulation. Moreover, it is hard to speculate on the biological role of Cas2 in systems that have been shown to target and incorporate DNA-derived spacers, although simultaneous integration of mRNA-derived protospacers or inhibition of invader proliferation by transcript degradation cannot be ruled out.

1.4.2 Stage II: CRISPR expression and biogenesis of crRNAs

The first observation that CRISPR loci are transcribed came from high-throughput analyses of non-coding RNAs in the archaeons *Archaeoglobus fulgidus* and *Sulfolobus solfataricus* P2 (Tang *et al.* 2002, 2005). The size distribution of the transcripts ranged from a minimum length corresponding to the distance between two successive repeats in the CRISPR cluster to higher order multiples of this single repeat-spacer unit. The detected sequences corresponded to various positions of the CRISPR arrays, implying that the whole loci are transcribed as long precursors (pre-crRNA), which are subsequently processed into smaller repeat-spacer units. Transcription of CRISPR loci has since been shown in a number of species, such as *E. coli* (Brouns *et al.* 2008), *Pyrococcus furiosus* (Hale *et al.* 2009), *Staphylococcus epidermidis* (Marrafini *et al.* 2008), *S. solfataricus* and *S. acidocaldarius* (Tang *et al.* 2005; Lillestol *et al.* 2009), *Xanthomonas oryzae* (Semenova *et al.* 2009). A constant

observation in all studies is the unidirectional transcription from the leader proximal end of the loci in all but one of the species studied so far, indicating the existence of a promoter within the leader region. Indeed, analysis of the transcription start sites and leader regions of the *Sulfolobales* revealed putative BRE and TATA box motifs within 25 nt of the transcription start side in the leader sequence (Lillestol *et al.* 2009). Reverse transcripts of the repeat clusters have only been detected in *S. solfataricus* and *S. acidocaldarius* by Lillestol *et al.* (2009), but their processing seems to be less efficient and therefore it remains unknown whether they produce functional repeat-spacer units. The authors attribute the production of the reverse transcripts to the existence of putative BRE and TATA box elements downstream of the CRISPR arrays, but whether this is a universal characteristic or purely coincidental remains unexplained, as does the functional relevance of the reverse crRNAs.

1.4.2.1 Regulation of CRISPR transcription

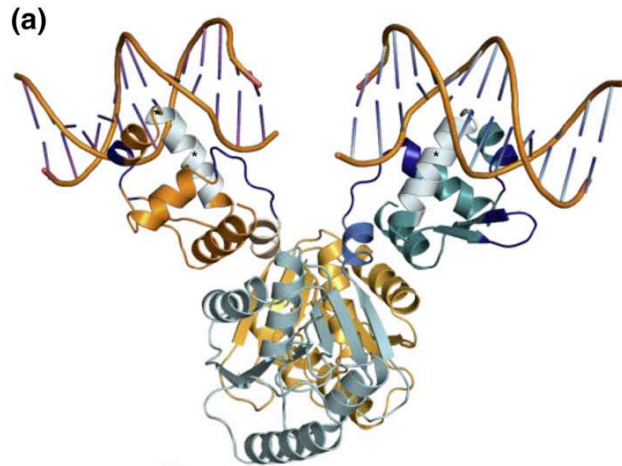
The factors that control the transcription of the CRISPR arrays and also the transcription and translation of Cas genes are still poorly understood. This procedure appears to be strikingly different between various CRISPR systems studied until now, which further highlights the remarkable versatility and ability of CRISPR systems to adapt and evolve according to environmental pressures. Transcription of CRISPR arrays and Cas genes have been shown to be either:

- i) constitutive, regardless of whether there is an ongoing infection or the state of it, consistent with a surveillance mode of action; interestingly, this is the case in all archaea studied to date (Tang *et al.* 2002; Hale *et al.* 2009; Semanova *et al.* 2009; Lillestol *et al.* 2006, 2009).
- ii) upregulated in response to phage infection, under control of the cAMP receptor protein (Agari *et al.* 2010). This pathway is also activated during carbon limitation stress. A recent study suggests that cas gene expression is also upregulated in response to envelope stress (Perez-Rodriguez *et al.* 2011)
- iii) subject to negative regulation by DevS along with the *dev* operon controlling developmental stages in *Myxococcus xanthus* (Viswanathan *et al.* 2007)
- iv) regulated by the antagonists H-NS and LeuO in *E. coli*. In this case, transcription under normal laboratory growth conditions is inhibited by the Heat-stable Nucleoid Structuring protein (H-NS, a typical transcriptional repressor in Gram-negative bacteria), which is bound to the promoter regions of the CRISPR locus and the Cas operon (Pul *et al.* 2019). This repression is relieved by the transcriptional regulator LeuO, by binding to the same genomic region and reversing the cooperative binding of H-NS dimers along the DNA, and also by directly or indirectly causing the enhancement of CRISPR-associated transcription (Westra *et al.* 2010).

Additionally, two novel putative transcriptional regulator families have been described recently in Archaea in the forms of *csa3* (*casRa*, COG0640/TIGR01884) and *csx1* (COG1517) (Lintner *et al.* 2011). Their activity and potential binding sites are still unknown/undetected, but the structure of Csa3 from *S. solfataricus* reveals a conserved binding site for a still unidentified allosteric small effector molecule, such as a dinucleoside polyphosphate (figure 1.10).

Figure 1.10: Structure of putative transcriptional regulator Csa3 from *S. solfataricus*

The active conformation is a homodimer. Chains are colored in teal and yellow/brown, and docked to dsDNA (adapted from Lintner *et al.* 2011).



1.4.2.2 CRISPR transcript processing

In contrast to the first stage of CRISPR functioning, processing of the CRISPR RNA transcript has been clarified for the most part in all three major CRISPR systems. This procedure is mediated by a single Cas processing endonuclease in each system, and at least in one case host RNAses have been shown to participate. Two specific functions have to be carried out by these enzymes, the first being the recognition of the precursor transcript and cleavage within a single site in each repeat to generate the mature form of crRNAs (figure 1.11), and the second the retention of the processed mature crRNA for subsequent usage by the respective effector proteins or complexes that mediate interference.

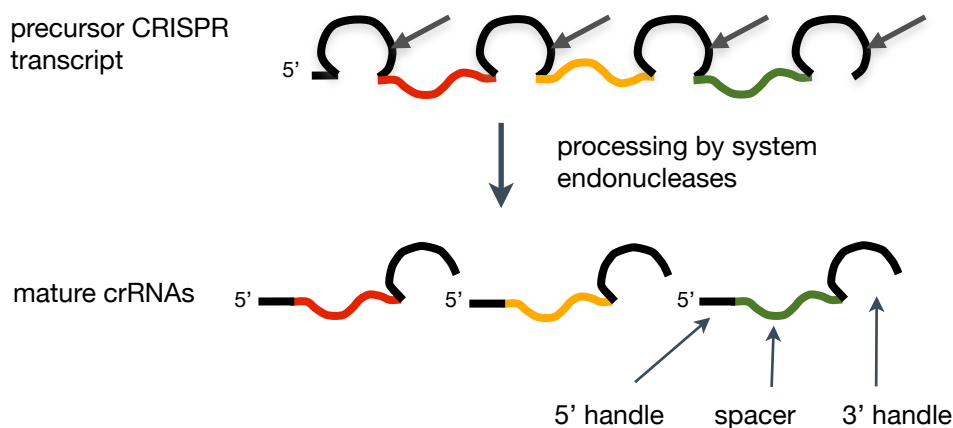


Figure 1.11: Outline of the second stage of CRISPR functioning

In CRISPR/Cas types I & III, a single superfamily of endonucleases, namely Cas6, are responsible for the processing of the primary transcript of the CRISPR locus into mature crRNA units that include a complete spacer flanked by parts of the repeat sequence (Carte *et al.* 2008, 2010). Cas6 family members are associated with subtypes I-A, I-B, I-D, III-A and III-B, while different families are found in subtypes I-E (names used in the literature: CasE/Cse3/Cas6e) and I-F (Csy4/Cas6f). These proteins are part of the RAMP superfamily (Repeat-Associated Mysterious Proteins) and have been shown to contain tandem or single ferredoxin-like folds, which contain the RRM motifs used to bind the target pre-crRNA (Carte *et al.* 2010; Haurwitz *et al.* 2010; Wang *et al.* 2011). Despite their shared fold and structural topology, the distinct families associated with each subtype exhibit remarkably different mechanisms for target RNA recognition and cleavage, although the final product is similar. This functional versatility is related to the specific repeat family of each subtype, as identified by Kunin *et al.* in 2007, as the propensity of each repeat sequence to form stable secondary structures (typically a stem-loop structure, depending on the palindromic nature of the repeat sequence) influences its mode of recognition and binding by the respective Cas proteins. Representatives of the three Cas6 families have been characterised biochemically and structurally, and their mode of action will be briefly described here.

In types I-E and I-F systems, the processing endonucleases (Cse3 and Csy4 respectively) are also subunits of the large multiprotein effector complexes that mediate interference. The first identified complex of this type was characterised in *E. coli* (type I-E) and termed CRISPR-Associated Complex for Antiviral Defence (acronym: CASCADE) (Brouns *et al.* 2008; Wiedenheft *et al.* 2011). The repeat sequences associated with this system is predicted to form a stable hexanucleotide stem with a tetranucleotide loop. The structure of Cse3 from *T. thermophilus* (Gesner *et al.* 2011; Sashital *et al.* 2011) is composed of a double ferredoxin-like fold, with a four strand antiparallel β -sheet forming the central positively charged cleft of the protein, where the phosphate backbone of the 3' strand of the stem loop is bound. Upon binding to RNA, the protein undergoes a conformational change whereby a previously disordered accessory β -hairpin recognizes the major groove of the RNA helix, and a previously disordered loop interacts with the base of the stem loop, positioning the scissile phosphate in the active site (figure 1.12 A). The protein interacts specifically with four residues located either side of the stem loop. Cleavage occurs at a G-A bond at the 3' base of the stem-loop. Mature crRNAs in this system, as sequenced from *E. coli* during the characterisation of CASCADE, comprise of a complete spacer sequence flanked by 8 nt of repeat derived sequence at the 5' end and the remaining 21 nt of repeat containing the stem-loop on the 3' end. A degree of heterogeneity was observed for the 3' end, highlighting the importance of the 5'

handle (or 5' psi-tag in the literature) for potential protein recognition and potentially in self-nonsel discrimination (Brouns *et al.* 2008; Jore *et al.* 2011). In type I-F systems, the C-terminal domain of Csy4 adopts an extended conformation although the basic secondary structure connectivity again resembles a ferredoxin-like fold (Haurwitz *et al.* 2010). The N-terminal domain is a typical ferredoxin-like fold. The stem-loop structure of the repeat interacts extensively with an arginine-rich helix in the C-terminal domain, while the ssRNA-dsRNA junction is positioned in the positively charged cleft between the two domains (figure 1.12 B).

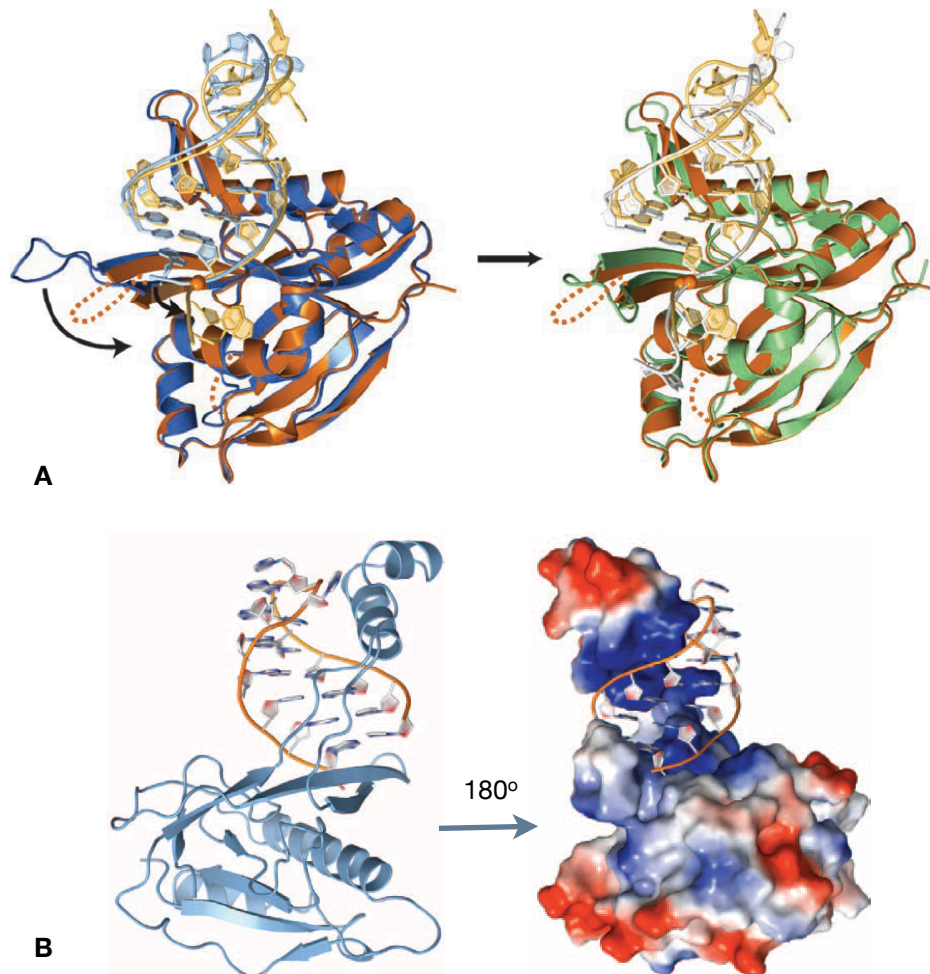


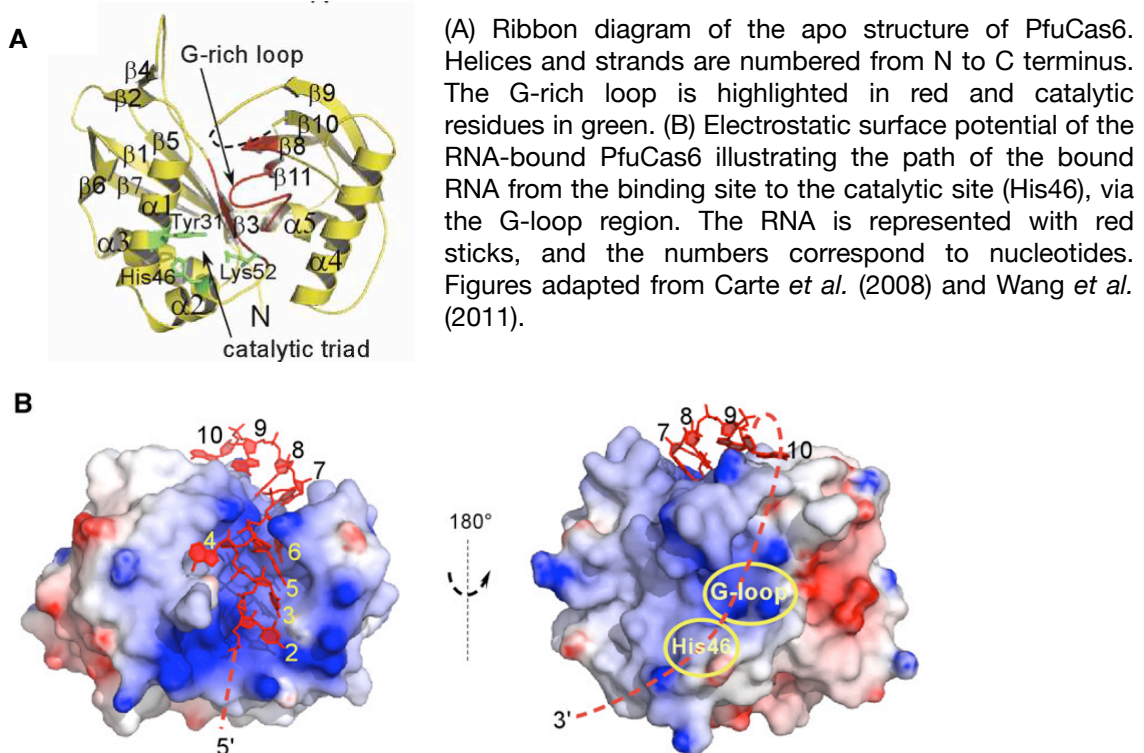
Figure 1.12: Structures of processing endonucleases Cse3 and Csy4

(A) Superimposition of two *T. thermophilus* Cse3 structures (in orange and blue) bound to synthetic CRISPR repeat RNA. The arrow indicates the conformational change occurring upon RNA binding. RNA is illustrated as a light orange tube, while the scissile phosphate as an orange sphere (adapted from Sashital *et al.* 2011). (B) Ribbon diagram and electrostatic surface representation of the structure of Csy4 from *P. aeruginosa* bound to the RNA CRISPR repeat substrate. The RNA backbone is represented with orange sticks. Blue shaded areas indicate the positively charged and red areas the negatively charged regions. Adapted from Haurwitz *et al.* (2010).

Sequence-specific hydrogen bonds tether the substrate in the active site so that the cleavage takes place immediately downstream of the hairpin, 8 nucleotides upstream of the spacer sequence. Both proteins remain bound to the cleavage products via the base-specific and electrostatic interactions formed with the RNA, which enables the subsequent use of the mature crRNAs by CASCADE and the analogous Csy complex.

A representative of the Cas6 family protein associated with subtypes I-A, I-B, I-D, III-A and III-B has been characterised in *Pyrococcus furiosus* (Carte *et al.* 2008, 2010; Wang *et al.* 2011). Although the architecture of this protein also consists of two ferredoxin-like domains it is apparent that the molecular mechanism for recognition and cleavage of the pre-crRNA has evolved to accommodate the type of unstructured repeat that is predicted to associate with these subtypes (Kunin *et al.* 2007). The conserved positively charged central cleft between the two ferredoxin-like domains is responsible for interaction with ssRNA, where conserved residues form contacts with specific conserved nucleotides near the 5' terminus of the CRISPR repeat, anchoring the RNA in position for the cleavage reaction taking place on the opposite surface of the protein (figure 1.13). Mutation analysis confirmed that the catalytic active site and binding site are physically distinct, with the connecting substrate interacting weakly or transiently with the signature Gly-rich loop. Metal-independent cleavage of the pre-crRNA transcript occurs 8 nt upstream of each spacer, producing the conserved 5' handle (termed psi-tag) present in the mature crRNA form and the 22 nt repeat-derived sequence at the 3' end.

Figure 1.13: Crustal structure of PfuCas6



The product remains bound to Cas6 until transferred to the respective effector complex (Cmr complex or an archaeal version of CASCADE, in the case of *P. furiosus* which contains both type I and III systems). The 3' end of the mature crRNA in *P. furiosus in vivo* is processed further by an unknown nuclease, but this seems to vary in different organisms (e.g. *S. solfataricus*). Cas6 family proteins have not been found to associate tightly with any effector Cas protein or complex, which grants them the flexibility needed to associate with multiple subtypes that potentially differ at the interference stage.

The catalytic mechanism used by all three types of processing endonucleases seems to rely on a histidine and a tyrosine residue in the active site, along with a variable lysine or serine, all of which are necessary for acid-base catalysis. Moreover, the glycine-rich loop characteristic of RAMP proteins is potentially implicated in correct substrate orientation. However, all three proteins use distinct sequence and structure-specific recognition mechanisms to select their respective substrates, illustrating the versatility of the characteristic duplicate ferredoxin-like fold in RAMPs and providing a mechanistic illustration of the coevolution of CRISPR repeat sequences and Cas proteins (Shah *et al.* 2010). To summarize, biogenesis of mature crRNAs in type I & III systems proceeds through single cleavage events within the repeat sequences 8 nt upstream of the beginning of the spacer. The generated sequence therefore consists of three elements: i) the strictly conserved repeat-derived 5' handle, predicted to be responsible for recognition and binding by the CASCADE-like effector complexes and/or determine target recognition as a self-nonself discrimination mechanism (discussed later); ii) the spacer sequence, responsible for target recognition by basepairing; iii) a heterogeneous repeat-derived 3' end, with a size range from 0 to 22 nt (Brouns *et al.* 2008; Hale *et al.* 2009; Carte *et al.* 2008; Haurwitz *et al.* 2010; Lintner *et al.* 2011). The processing events that lead to trimming of the 3' end are still unidentified, as is the functional significance (if any) of this heterogeneity.

A quite remarkable procedure for CRISPR RNA maturation takes place in type II systems, as discovered in *Streptococcus pyogenes* by Delcheva *et al.* (2011). In this system, a novel RNA species was found in high copy number and identified as the transcript of the opposite strand of a region upstream from the start of the *cas* operon and the CRISPR array. Interestingly, a 25 nt region of this transcript, termed tracrRNA (trans-activating CRISPR RNA), was complementary to the repeat sequence of *S. pyogenes* (with only one mismatch), which is predicted to be unstructured. It was demonstrated that an RNA duplex formed by the tracrRNA and a repeat sequence in the pre-crRNA is sufficient to guide the cleavage of both strands at specific positions within the duplex region by the host RNase III, producing 1x crRNA units that consist of a complete spacer sequence flanked by the partial repeats. Further processing

takes place on the 5' end of the spacer sequence by a still unidentified nuclease, resulting in the mature crRNA form for this system (figure 1.14, Delcheva *et al.* 2011). The latter comprises solely of a 5' 20 nt spacer-derived sequence and a 19-22 nt repeat-derived sequence on the 3' end. This composition is strikingly different from the mature form of crRNAs found in types I and III in that it lacks the characteristic 5' repeat-derived handle. This feature could indicate a distinct mechanism for crRNA recognition by the proteins mediating the interference and potentially for the interference itself. The only Cas protein implicated in this stage is Cas9 (Csn1) although its exact function is unknown. In the model proposed by the authors the duplex formation between the tracrRNA and the pre-crRNA is enabled by Cas9, prior to recognition and cleavage of both strands by the host RNase III in a process termed trans RNA mediated activation of crRNA maturation (Delcheva *et al.* 2011). Cas9 contains a McrA/HNH-nuclease domain and a RuvC-like (RNase H-like) domain (Makarova *et al.* 2006), making it a suitable candidate for the second cleavage event. There is no indication that the role of Cas9 is restricted at this stage, as it is possible that it also participates in the interference mechanism as deletion of *cas9* in *S. thermophilus* resulted in loss of phage resistance (Barrangou *et al.* 2007). To this date, this is the first example of a host factor implicated in CRISPR function, highlighting the exceptional economy and versatility of this system.

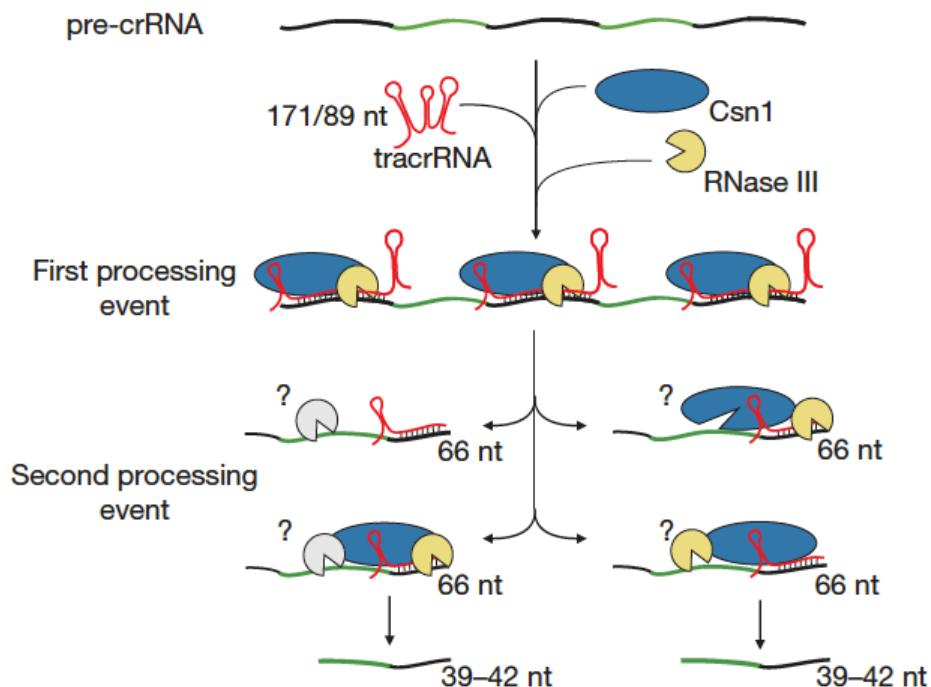


Figure 1.14: Model for CRISPR RNA processing in type II systems

In the first processing event, basepairing between the tracrRNA and the repeats (black) in the precursor CRISPR transcript (spacers are in green), lead to site-specific cleavage by RNaseIII in the repeat sequence, generating repeat-spacer units. The second, still unidentified processing event takes place within the spacer sequence, generating the mature crRNA units in type II systems. (adapted from Delcheva *et al.* 2011)

1.4.3 Stage III: Recognition of invader sequences and target interference

The identity of the molecular target of the CRISPR system has been a subject of debate since the discovery of its role in defence against extrachromosomal elements. The first bioinformatics studies providing an initial thorough description and classification of the system's protein and genetic components (Jansen *et al.* 2002, Haft *et al.* 2005, Makarova *et al.* 2006) detected a functional analogy with the eukaryotic siRNA-interference system and hypothesised that RNA would be targeted. Subsequent studies overturned this hypothesis, as a closer inspection of the distribution of protospacers in phage, viral and plasmid genomes revealed that both sense and antisense strands were represented in CRISPR spacers, and no bias to gene-coding regions or conserved genes was observed (Shah *et al.* 2009). This result had two implications: firstly it suggested that the source of the spacers is the invader DNA itself in a random and non-directional manner, and not the viral or plasmid mRNA transcripts, and secondly it indicated that interference could also occur at the DNA level and not at a gene expression level (Shah *et al.* 2009). Additionally, if RNA was indeed the target, an over-representation of phage and viral genes that are expressed early in the lytic cycle would be expected as a more efficient inhibitory mechanism, which is not the case. This random unified distribution was demonstrated for the CRISPR spacers, virus families and plasmids of the *Sulfolobales* (Lillestol *et al.* 2006; Shah *et al.* 2009), other crenarchaeal neutrothermophiles such as *Aeropyrum pernix* and *Pyrobaculum sp.* (Shah *et al.* 2009), phages and plasmids of *Streptococcus thermophilus* (Barrangou *et al.* 2007; Deveau *et al.* 2008; Horvath *et al.* 2008), prophages and non-viral regions of the *Yersinia pestis* chromosome (Cui *et al.* 2008) and plant pathogen *X. oryzae* (Semenova *et al.* 2009). Moreover, as mentioned in the previous paragraph transcription of the CRISPR loci in most organisms has been shown to be unidirectional, suggesting that the generated crRNAs must be able to recognise their complementary targets regardless on whether they are located on the sense or antisense strand. Finally, it has to be noted that the current characterisation of RNA viruses is poor, therefore it is unsurprising that no spacer matches to RNA bacteriophages have been identified yet for bacterial CRISPR systems, and no RNA archaeal viruses have been isolated.

Solid experimental evidence regarding this matter has been provided in four bacterial systems to date, namely *E. coli*, *Staphylococcus epidermidis*, *Streptococcus thermophilus* and *Pseudomonas aeruginosa*. *E. coli* carries a type I-E CRISPR/Cas system, comprising of eight cas genes (figure 1.15), the core cas1-3 and subtype specific cse1-4 and cse5e, and a single CRISPR locus. Brouns *et al.* (2008) demonstrated that a large multimeric complex composed of Cse1-5e (alternatively named CasA-CasE) can be co-purified from *E. coli* lysate by affinity chromatography,

and termed it CASCADE (Crispr Associated Complex for Antiviral Defence). The complex could process transcribed crRNA from *E. coli* and bind the mature crRNA units, which in *E. coli* consist of a complete spacer with the last eight nucleotides of the repeat sequence on the 5' end (5' psitag) and a less defined 3' end with the remaining nucleotides of the repeat (3' handle). Construction of artificial crRNAs against both the coding and template strands of phage lambda resulted in inhibition of virus proliferation when CASCADE and Cas3 were present, providing the first direct evidence that DNA is the target of interference. In this context, the complex-bound crRNA serves as a guide to identify invading DNA and recruit the effector molecule, in this case Cas3, that will complete the silencing procedure presumably by degradation of the target sequence. A functionally analogous complex was isolated from the type I-F system of *P. aeruginosa*, comprising of subunits *csy1-4* (Wiedenheft *et al.* 2011).

In the second case, a clinical isolate of *S. epidermidis* was found to carry a type III-A CRISPR/Cas system consisting of core genes *cas1*, *cas2*, *cas6*, subtype specific genes *csm1-6* and a single CRISPR locus (figure 1.15). Staphylococcal conjugative plasmids contain a protospacer match within the conserved nickase gene, the transcription of which in the recipient cell is not essential for conjugation. Marraffini *et al.* (2008) demonstrated that conjugative transfer of the plasmid was inhibited in the strain that contained the CRISPR system, providing a first clue that DNA is targeted in this system. Insertion of a self-splicing intron into the centre of the protospacer resulted in the disruption of the original target at the DNA level, but after conjugation, transcription of the respective gene and splicing of the mRNA, the target would be reconstituted at the RNA level. The plasmid was able to propagate efficiently, indicating that the target mRNA was not recognised by the CRISPR system. Additional experiments in which plasmid transformation was inhibited due to interference against matching protospacers regardless of their orientation in the plasmid or their active transcription, confirmed that invader DNA is the original target regardless of its source (plasmid or virus) or the transfer mechanism (infection, conjugation or transformation).

Finally, it was established directly that the CRISPR/Cas type II system of *Streptococcus thermophilus* is able to cleave both bacteriophage and plasmid DNA *in vivo*. The CAS gene cluster associated with CRISPR locus 1 in *S. thermophilus* consists of genes *cas9*, *cas1*, *cas2*, *csn2* (figure 1.15) and addition of novel spacers after exposure to foreign genetic elements such as plasmids or bacteriophages was observed only in this locus out of the four found in this organism. Garneau *et al.* (2010) managed to isolate linearised plasmids from adapted strains exhibiting only partial interference, and map the cleavage site in the protospacer sequence, 3 bp upstream of the protospacer adjacent motif (PAM: conserved tri- or dinucleotide motif found at the 3' end of protospacers, see 1.4.4). The same result was observed in

bacteriophage DNA extracted from infected strains carrying appropriate CRISPR spacers. The location of the cleavage sites in the phage genome was identical to those observed in the linearised plasmid. Additionally, a second cleavage site was detected in protospacers in the positive strand of the phage genome, 19 or 20 bases upstream of the PAM. This suggests a measuring mechanism anchored in the 3' end of the protospacer (where the PAM is located), and is reminiscent of the crRNA-guided cleavage of the RNA target in *Pyrococcus furiosus* (Hale *et al.* 2009). In sensitive strains or in strains where *cas9* was deleted only the circular form of the plasmid or the intact bacteriophage genome was detected, confirming the previously suggested involvement of Cas9 in the interference stage (Barrangou *et al.* 2007). It should be noted that inactivation of *Csn2* inhibited the insertion of new spacers and the generation of new resistance mutants, confirming its involvement in the adaptation stage. Thus it is demonstrated that the interference-mediating protein in this system (possibly Cas9, since it contains an HNH nuclease domain) exhibits endonuclease activity against foreign dsDNA (either plasmid or phage) using a molecular ruler mechanism guided potentially by the crRNA, and producing blunt ends. The orientation-dependent differential cleavage pattern between protospacers located in the sense or antisense strand of the invading element remains unexplained. The number of cleavage sites in the phage genome corresponded to the number of protospacer matches, in accordance with the authors' previous observation that the number of acquired spacers against a particular invading element has a cumulative effect on the resistance displayed against the respective element. The mechanistic details of this system are still unknown, therefore we do not know whether the target identification proceeds via the formation of an R-loop as exhibited for the CRISPR I-E system in *E. coli*.

The first archaeal system in which CRISPR/Cas mediated defence was demonstrated *in vivo* was the crenarchaeon *Sulfolobus*, in particular strains *S. solfataricus* P2 (Manica *et al.* 2011, Gudbergsdottir *et al.* 2011) and *S. islandicus* REY15A (Gudbergsdottir *et al.* 2011). Types I-A and III-B CRISPR/Cas systems coexist in this archaeon, and the gene organisation of type I operons in *S. solfataricus* P2 can be seen in figure 1.15. Manica *et al.* exploited the fact that a natural spacer in CRISPR locus B of *S. solfataricus* P2 matches a gene-coding region of the conjugative plasmid pNOB8. Transformation efficiencies for a recombinant SSV1 virus carrying the aforementioned protospacer in *S. solfataricus* P2 strain were very low compared to strain *S. solfataricus* P1 (which does not carry the respective spacer) regardless of whether the protospacer was transcribed and despite the fact that it was not essential for the virus propagation. Additionally, recombinant SSV1 shuttle vectors carrying a mini-CRISPR locus with self-targeting spacers (against a chromosomal gene) in both orientations were unstable in the host *S. solfataricus* P2 strain and underwent

recombination in order to eliminate the self-targeting spacer. Similarly, Gudbergsdottir *et al.* observed that when challenged with shuttle vectors containing viral genes with CRISPR-matching protospacers, very few transformants were able to survive only after deleting part of the chromosomal CRISPR locus including the relevant spacer, or whole CRISPR/Cas modules, regardless of whether the protospacers were transcribed. Both these observations indicate that interference occurs at a DNA level, although it was not resolved which Cas type was responsible for this phenotype.

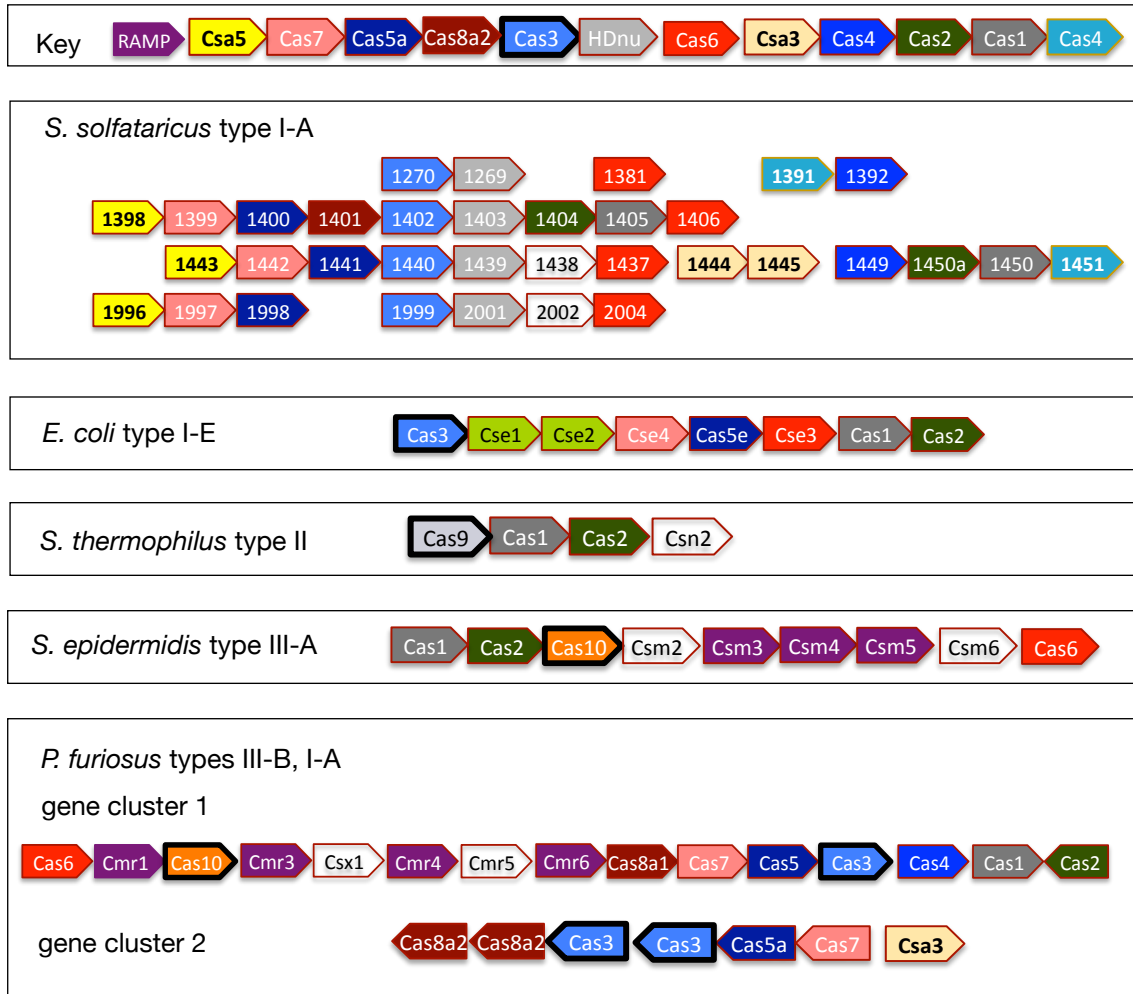


Figure 1.15: Gene organisation of Cas operons of studied organisms

Cas gene neighbourhoods of *S. solfataricus* (CMR operon not illustrated here), *E. coli*, *S. thermophilus*, *S. epidermidis* and *P. furiosus*, organisms for which CRISPR functioning has been characterised experimentally. Numbers in arrows represent ORFs. Colour-coding indicates homology, except between blank arrows which indicate subtype-specific or non-cas genes. Orientation of the arrowheads indicates direction of transcription. Arrows outlined in black indicate signature genes for the specific subtype. Nomenclature according to Makarova *et al.* 2011.

Conveniently, all the cases described above cover all three major CRISPR/Cas types (although not all the subtypes), suggesting that crRNA-directed DNA recognition and cleavage is a general mechanism for CRISPR mediated interference. Considering the complexity and diversity of the system components and their organisation, it is obvious that general predictions cannot be made and multiple pathways will emerge, as this diversity is also reflected at a biochemical level. However, from the host cells' point of view the physiological importance of interference occurring at the DNA level is enormous, as it would dramatically increase the target and temporal activity range of the defence system.

The exact mechanism by which Cas-crRNA effector complexes recognise and access their target sequence within a given DNA genome has recently been determined by studies with the multiprotein CASCADE-like complexes in *E. coli* and *P. aeruginosa* (Wiedenheft *et al.* 2011; Semenova *et al.* 2011). In both systems the target is recognised by base-pairing to an 8 nt or 7 nt "seed" sequence located at the 5' end of the spacer sequence in the crRNA (figures 1.16, 1.17). Affinity for the rest of the spacer sequence was shown to be much smaller, accounting for the tolerance for protospacer mismatches observed in many cases (Deveau *et al.* 2008; Gudbergsdottir *et al.* 2011 among others).

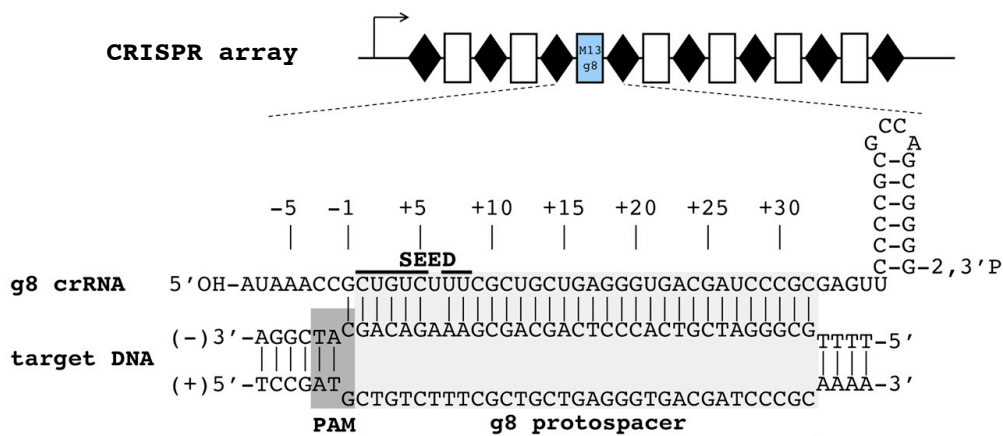


Figure 1.16: Basepairing between crRNA and protospacer upon target recognition in *E. coli*. The high-affinity seed sequence consists of 7 nt (non-contiguous) on the 3' end of the protospacer, adjacent to the PAM. Numbering starts from the first nucleotide of the crRNA spacer. In the CRISPR array, repeats are drawn as rhombi, spacers as rectangles (adapted from Semenova *et al.* 2011).

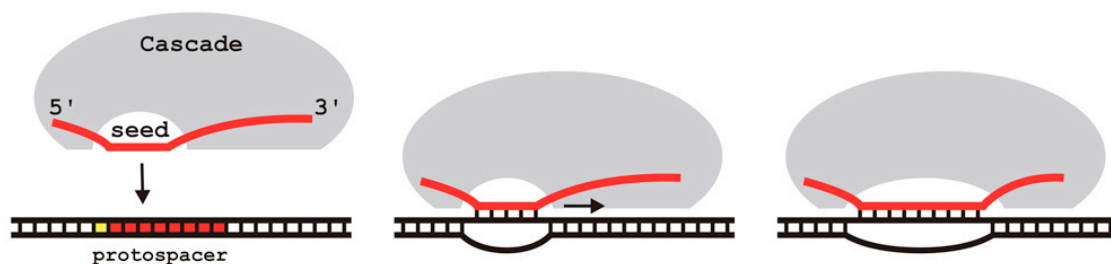


Figure 1.17: Model for target recognition by CASCADE

The search and initial recognition of a protospacer is mediated by its seed sequence. Initial binding promotes further hybridisation between the full crRNA spacer sequence and the protospacer, leading to localised duplex unwinding (adapted from Semenova *et al.* 2011).

The only example of RNA targeting by a CRISPR/Cas system comes from *Pyrococcus furiosus* (Hale *et al.* 2009). As described in detail in Chapter 3, the type III-B Cmr complex (acronym standing for CRISPR module RAMP) (figure 1.15) from this hyperthermophilic euryarchaeon exhibits cr-RNA guided endoribonuclease activity against ssRNA targets *in vitro*. This type of interference *in vivo* has not yet been demonstrated, and the spacer sequences from *P. furiosus* do not contain matches to any known viruses or plasmids, therefore the questions raised about the biological function of such an activity cannot be answered.

1.4.4 Protospacer selection, self/non-self discrimination and autoimmunity issues

The subjects of protospacer selection from the invader genome for incorporation into CRISPR loci and discrimination between the exogenous and the host DNA are two of the most important unanswered questions in CRISPR functioning. An increasing number of studies suggest that short sequence motifs adjacent to the protospacer are implicated in both processes, albeit with a still unknown mechanism.

Deveau *et al.* (2008) and Horvath *et al.* (2008) first identified a highly conserved tetranucleotide motif located two nucleotides downstream of phage protospacer sequences matching CRISPR spacers in *S. thermophilus*. Mutations in this motif enabled phages to escape CRISPR immunity, suggesting a role in target recognition. These motifs were identified as a universal CRISPR characteristic by Mojica *et al.* (2009), who found that strict motif conservation was limited to 2-3 nt located one position after the 3' end of the protospacer and introduced the term Protospacer Adjacent Motifs (PAMs). The consensus motif sequences depend on the CRISPR repeat group, as assigned by Kunin *et al.* (2007). Moreover, the PAMs appear to determine the spacer orientation in relation to the protospacer, as the spacer end that corresponds to the PAM - proximal side of the protospacer is always oriented towards

the leader sequence (figure 1.18, Mojica *et al.* 2009). This conserved orientation along with the fact that novel spacers are always added at the leader-proximal end of a CRISPR locus indicates a potential role in the selection of protospacer sequences and integration procedure, presumably as a binding sequence for Cas proteins (Mojica *et al.* 2009).

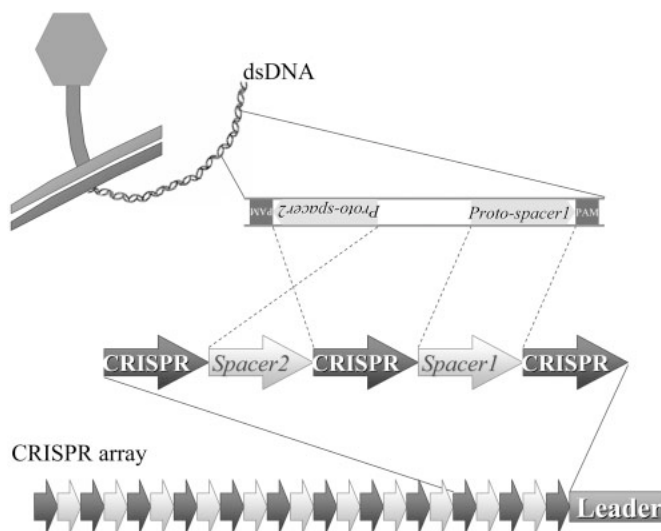


Figure 1.18: Orientation of protospacers in regard to their PAM. Protospacers 1 and 2 are located in opposite strands, but are always incorporated with the PAM-proximal side towards the leader (adapted from Mojica *et al.* 2009).

Subsequent studies confirmed the role of the PAMs in interference efficiency, as any mutation within the motif prevented CRISPR interference and led to successful infection of the virus/plasmid carrying the respective mutation, while an intact PAM motif was required for interference (Semenova *et al.* 2009; Marraffini and Sontheimer, 2010; Gudbergsdottir *et al.* 2010). However, studies in *P. furiosus* (type III-A) and *S. solfataricus* (types I-A and III-B) did not detect a role for the PAM in CRISPR interference (Hale *et al.* 2009; Manica *et al.* 2011), indicating that the importance of this motif is still elusive.

The mechanism by which the CRISPR system distinguishes the cognate CRISPR spacers from the invader protospacer sequences (both of which would exhibit perfect complementarity to the respective crRNA and would therefore constitute interference targets in the absence of such a mechanism) was elucidated by Marraffini and Sontheimer (2010) using engineered conjugative plasmids of *Staphylococcus epidermidis* (type III-A). The authors demonstrated that the differential complementarity between the generally conserved 5' handle of crRNA and the 3' region downstream of the target protospacer (that is, beyond the region of complete basepairing between the crRNA spacer and the target protospacer) is responsible for discrimination between self and non-self sequences and subsequent interference. This region is fully complementary only to the endogenous CRISPR repeat sequences, and if basepairing occurs (especially in positions -2, -3 and -4 of the 5' handle), then the target DNA is protected. If no complementarity is found between the 5' handle and

the downstream region of the protospacer, then the system proceeds with target cleavage (figure 1.19). Although the region screened for complementarity on the invader DNA corresponds to the PAM location, involvement of a specific motif was not observed in this mechanism, supporting the putative role of PAM in protospacer selection.

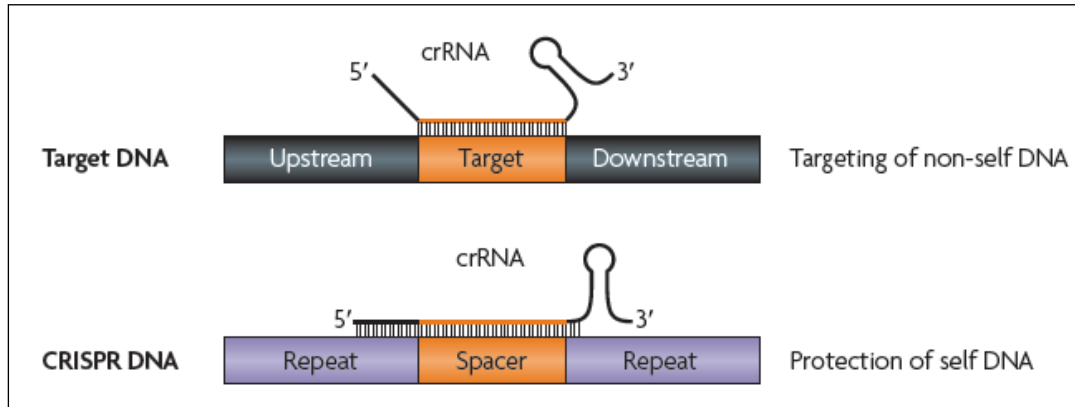


Figure 1.19: Model for discriminating between self and non-self DNA during CRISPR target recognition

Complementarity between the repeat-derived 5' and 3' handles of the crRNA and the target DNA ensures protection of the endogenous CRISPR locus (adapted from Marraffini and Sontheimer, 2010).

The deleterious consequences of autoimmunity are well known for all existing immunity systems, whether prokaryotic or eukaryotic. In the CRISPR/Cas system, this risk is manifested with the incorporation of a spacer into a CRISPR array that targets a cognate sequence. Although not abundant, self-targeting spacers have been detected in CRISPR loci (Mojica *et al.* 2005; Bolotin *et al.* 2005; Horvath *et al.* 2008, 2009; Shah *et al.* 2009), leading to the suggestion that, in analogy to the eukaryotic RNAi, CRISPR arrays could also have a regulatory role within the organism (van der Oost *et al.* 2009). An extensive case analysis by Stern *et al.* (2010) disproved this theory by demonstrating that, although self-targeting spacers were present in 18% of the CRISPR-encoding strains analysed in this study (330 total), they do not exhibit conservation across species (as would be expected for a successful regulatory element) and are always accompanied by one of the following adaptations: deletion of the respective spacer, loss of the Cas operon, inactivation of the CRISPR locus, degeneration of the flanking repeats, mutation of the target self-protospacer, or inferred deletion of the whole CRISPR array (Stern *et al.* 2010). These observations confirm that the accidental incorporation of a self-targeting spacer leads to negative autoimmunity effects potentially lethal for the host, and various adaptations have been observed to ensure inactivation of the specific spacer.

This procedure was documented *in vivo* by Manica *et al.* (2011), when a plasmid containing an engineered small CRISPR locus carrying spacers against an

endogenous host gene was introduced in *S. solfataricus*. The attack was lethal, and the few surviving cells had either lost the plasmid, or recombination had occurred between the plasmid and the genomic CRISPR loci in order to replace the self-targeting spacer.

1.5 Evolution, mobility and distribution of the CRISPR/Cas system

Phylogenetic analysis of *cas* genes and the presence of CRISPR loci on conjugative plasmids has led to the conclusion that CRISPR/Cas, like previously described prokaryotic defence systems, is a mobile genetic element dispersed among species by horizontal gene transfer (Haft *et al.* 2005; Makarova *et al.* 2006; Godde *et al.* 2006).

CRISPR loci also exhibit a great degree of plasticity and rapid evolution rates, as can be observed by the hypervariability of CRISPR regions in terms of repeat/spacer content between closely related strains or between generations (Lillestol *et al.* 2009; Andersson and Banfield, 2008; He and Deem, 2010). For example, studies in natural microbial communities have shown that CRISPR regions evolved rapidly enough by incorporating/losing spacer sequences in response to environmental pressure to promote cell individuality, therefore ensuring population immunity (Tyson and Banfield, 2008; Andersson and Banfield, 2008). Large deletions produced by low levels of spontaneous internal recombination of CRISPR loci are also common, and it is thought to be a way of limiting the ever-increasing size of the arrays and the energetic cost to the organism (Lillestol *et al.* 2009; Gudbergdottir *et al.* 2010).

Makarova *et al.* (2011a) attempted to resolve the evolutionary and phylogenetic relationships between the CRISPR/Cas systems by combining phylogenetic information from Cas1 homologues (being the universal CRISPR signature gene), genomic context analysis of the Cas operons and correlation with the CRISPR repeat groups (as defined by Kunin *et al.* 2007). The resulting dendrogram is illustrated in the left part of figure 1.7.

CRISPR/Cas systems exhibit a differential distribution among the main archaeal and bacterial taxa. In general, the system is overrepresented in the kingdom of Archaea, with almost 90% presence in the sequenced genomes compared to about 40% in bacterial genomes (table 1.1, Makarova *et al.* 2011a). Moreover, type III systems are more often associated with archaea, in particular thermophilic species, while type II systems seem to solely present in bacteria (table 1.1, Makarova *et al.* 2011a). Also, archaea tend to carry more than one unrelated CRISPR/Cas systems, occupying up to 1% of their chromosome. Drawing from these observations, it has been proposed by Makarova *et al.* (2011a and b) that the system originated in an elementary form in thermophilic archaea, and subsequently spread by HGT.

1.6 CRISPR/Cas and the eukaryotic RNAi

RNA interference (RNAi) is a method of sequence-specific gene silencing mediated by a variety of non-coding, small RNA species with the aid of specific protein complexes in eukaryotes. Different pathways of this mechanism involve RNA species of different origin and serve different functions (reviewed in Hannon, 2002; Meister and Tuschl, 2004; Malone and Hannon, 2009; Carthew and Sontheimer, 2009): i) genome-encoded microRNAs are involved in post-transcriptional regulation of gene expression in the miRNA pathway; ii) exogenous dsRNA is processed into short interfering RNAs (siRNAs) which are then used to silence the invading element; iii) endogenous piwi-interfering RNAs are used to silence transposons in animal germ cell lines. This by no means represents an exhaustive list, as novel classes of small RNA species with distinct functions are still being discovered, while the stages of small RNA biogenesis and target silencing are not yet fully elucidated. A brief summary of the general RNAi pathway will be presented here (as reviewed in Carthew and Sontheimer, 2009; Siomi and Siomi, 2009), in order to discuss the proposed analogies with the CRISPR system.

In the first stage, long dsRNA molecules of exogenous origin are processed into 21-25 nt duplex RNA fragments with a 3' overhang (siRNAs) by an RNase III endonuclease called Dicer. The cleavage is non-sequence specific, but Dicer operates as a "molecular ruler" determining the size of the produced siRNAs. In contrast, the precursor molecules for miRNAs are transcribed from the organism's chromosome (often encoded within introns) and fold into an imperfect stem-loop structure, which is then processed into miRNAs 21-25 nt in length by Dicer. In the second stage, the siRNA is unwound and one strand (termed the guide strand) is selected for incorporation in the active version of a large ribonucleoprotein complex called RISC (RNA-Induced Silencing Complex). The composition of the RISC varies in different organisms, but the core component is always a member of the Argonaute protein family, also known as Slicer. This protein binds the guide ssRNA and uses it to locate complementary RNA strands, which are subsequently cleaved by a conserved domain of the Argonaute (PIWI domain). The cleavage site is determined again by a "molecular ruler" mechanism based on the size of the guide RNA. The RISC complex is then recycled and can repeat the silencing procedure. The outcome of the RISC encounter with the target RNA depends largely on the degree of complementarity exhibited between the guide and the target RNA, the type of Argonaute protein, the specific subunits of the RISC complex and other proteins interacting with the target and/or the RISC (summary in figure 1.20).

Functional analogies were detected initially by Makarova *et al.* (2006) between the then-hypothetical mode of CRISPR functioning and RNAi. Strategic similarities focused on the following facts: they both rely on an anti-sense RNA mechanism for sequence-specific target silencing; mature RNA molecules (crRNA and siRNA) derive

from long RNA precursors and contain invader-derived sequences; target RNA cleavage by a Cas multiprotein complex had been demonstrated *in vitro* (Hale *et al.* 2009); the abundance of conserved RNA and DNA-manipulating domains within the array of proteins associated with each system (reviewed in Marraffini and Sontheimer, 2010). This led to the hypothesis that CRISPR is homologous to RNAi and represents a prokaryotic version of the same system (Makarova *et al.* 2006). However, with the elucidation of the mechanistic details of the various stages of CRISPR functioning, this hypothesis was eventually discarded. First, the respective protein machineries and the small RNA biogenesis were proven to be completely different. Second, the ultimate target for CRISPR interference was shown to be DNA (with the exception of the RNA-targeting Cmr complex in *P. furiosus*; Hale *et al.* 2009). Third, each system seemed to have distinct physiological roles, with RNAi involved (among other functions) in gene regulation, chromosome stability, transposon and invader silencing, while the CRISPR system seems to be predominantly an immune system (reviewed in Marraffini and Sontheimer, 2010; Horvath and Barrangou, 2010).

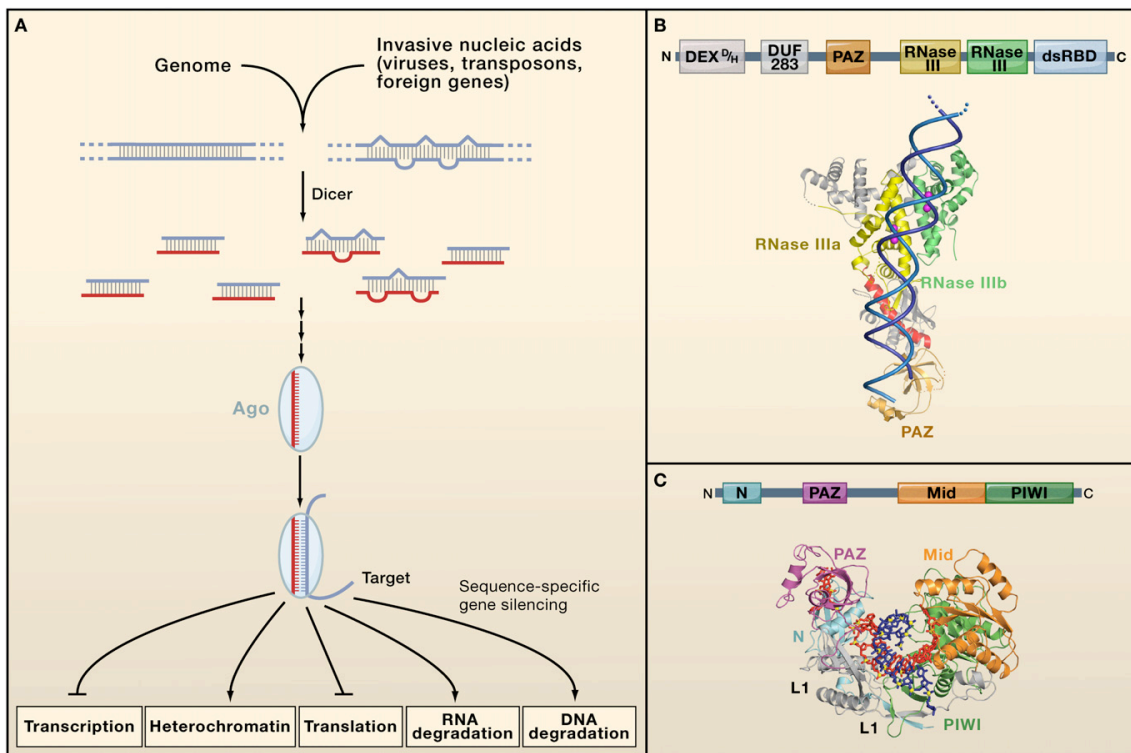


Figure 1.20: General pathway and key proteins for RNA interference

(A) The sources of long dsRNA precursors can be genomic or exogenous. Processing by Dicer generates the mature siRNAs or miRNAs. The guide strand (red) of each si- or miRNA is loaded onto the Argonaute (Ago) protein, and used to recognise the complementary target. The exact method of silencing depends the location and nature of the target. Domain organisation and structure of Dicer (B), loaded with dsRNA and the Argonaute (C). Adapted from Carthew and Sontheimer, 2009.

1.7 Applications and alternative roles for CRISPR/Cas

The extreme sequence variability and mechanistic differences observed between the CRISPR/Cas subtypes, along with the biased spacer content detected in certain cases (Lillestol *et al.* 2009) have raised the possibility that the various system subtypes might be functionally distinct (Kunin *et al.* 2007), or that the CRISPR/Cas system may be involved in other cellular processes apart from antiviral defence (Deveau *et al.* 2010; Karginov and Hannon, 2010). The presence of self-targeting spacers, although proven to be unsustainable and deleterious to the host by Stern *et al.* (2010) is still suggested as a putative function in certain cases (Shah *et al.* 2009, Touchon *et al.* 2010). CRISPR loci, due to their extensive repeat homology, were involved in large-scale genome rearrangements in Thermotogales (DeBoy *et al.* 2006). In *P. aeruginosa*, the CRISPR/Cas system was implicated in regulating group behaviors such as swarming motility and biofilm formation in response to lysogenisation by the DMS3 phage (Zeagans *et al.* 2009). These phenotypes were lost if the DMS3 phage infected a strain harboring a matching CRISPR spacer, indicating that this loss is potentially a resistance mechanism, as it would prevent the phage from spreading to the population by proximity. In *Myxococcus xanthus*, regulation of the cas operon is linked to the dev operon, a cluster involved in fruiting body development (Viswanathan *et al.* 2007). Purified Cas1 from *E. coli* was shown to interact with essential DNA repair enzymes such as RecB, RecC and RuvB (Babu *et al.* 2010). A potential functional association was supported when Cas1 or CRISPR loci deletion strains exhibited sensitivity to DNA damage and impaired cell division. Finally, recently the CRISPR system in *E. coli* was shown to be upregulated by factors involved in detecting and triggering downstream responses to envelope stress (Perez-Rodriguez *et al.* 2011). It should be noted here that this would not be the first example of a prokaryotic defence system involved in additional functions, especially stress response pathways (see paragraph 1.1). The adaptability and versatility of the prokaryotic resistance systems seems to be a general feature, but for the time being these examples only refer to individual cases and no function other than immunity has been widely associated with the CRISPR/Cas systems.

The current and potential applications of the CRISPR/Cas system are the following (as summarised by Sorek *et al.* 2007; Al-Attar *et al.* 2011):

- i) CRISPR regions are already used for strain identification using a method called spacer-oligotyping (spoligotyping) (Kamerbeek *et al.* 1997)
- ii) Engineering anti-viral systems for bacterial strains in dairy industry, targeted against desired phages.
- iii) Selective gene silencing in bacteria/archaea by exploiting the RNA interference mechanism, thus avoiding the laborious knockout procedure.

- iv) Exploit the unmatched spacer content to identify novel viruses in natural samples. (Snyder *et al.* 2010)

1.8 CRISPR/Cas systems of *Sulfolobus solfataricus*

The hyperthermophilic crenarchaeon *Sulfolobus solfataricus* has been a model organism for the physiological and biochemical studies of Archaea and their evolutionary relations to Bacteria and Eukaryotes, even before the sequencing and publishing of its genome (She *et al.* 2001). Cas systems in this organism belong to subtypes I-A and III-B with multiple operons for each subtype, the organisation of which and respective gene names can be seen in figure 1.21

S. solfataricus strain P2 contains seven CRISPR loci with a total of 425 repeats, the properties of which can be seen in table 1.2. Five of the loci have adjoining partially conserved leader sequences, (Lillestol *et al.* 2006), indicating that they are active. Indeed, transcripts could only be detected from the leader-carrying loci (Lillestol *et al.* 2009). Comparative analysis of the repeat sequences, leader motifs, spacer matches and proximal cas operons of CRISPR/Cas systems of the Sulfolobales enabled the classification of the CRISPR loci in three major families (I-III) and elucidation of their phylogenetic relationship (Lillestol *et al.* 2009). Distinct PAM motifs were identified in the matching protospacers for each family, namely CC for family I, TC for family II and GT for family III protospacers. The authors did not detect preferential incorporation of spacers from specific viruses/plasmids into any of the families, therefore it remains unknown whether they have a functional distinction. Analysis of the leader regions preceding each locus revealed the presence of low complexity conserved sequence motifs, which appear to correlate with the respective CRISPR family of the locus. The importance of these motifs is yet to be determined, although it is hypothesised that they consist of recognition elements for Cas protein binding, or serve other regulatory functions (Lillestol *et al.* 2009). *S. solfataricus* CRISPR loci belong to families I and II. Analysis of 415 spacers in *S. solfataricus* with an average length of 38-42 bp revealed similarities to sequences from Rudiviruses, β - Lipothrixviruses, Fuselloviruses, STIV, ATV, SIRV viruses, pRN family plasmids and the conjugative plasmid pNOB8 (Mojica *et al.* 2005; Shah *et al.* 2009). No self-targeting spacer was identified within the total of 135 matched spacers. The repeat sequences in the CRISPR loci of *S. solfataricus* are not predicted to form any significant secondary structures, as they belong to a repeat cluster with extremely low folding scores and therefore predicted to be unfolded (cluster 7), as classified by Kunin *et al.* (2007).

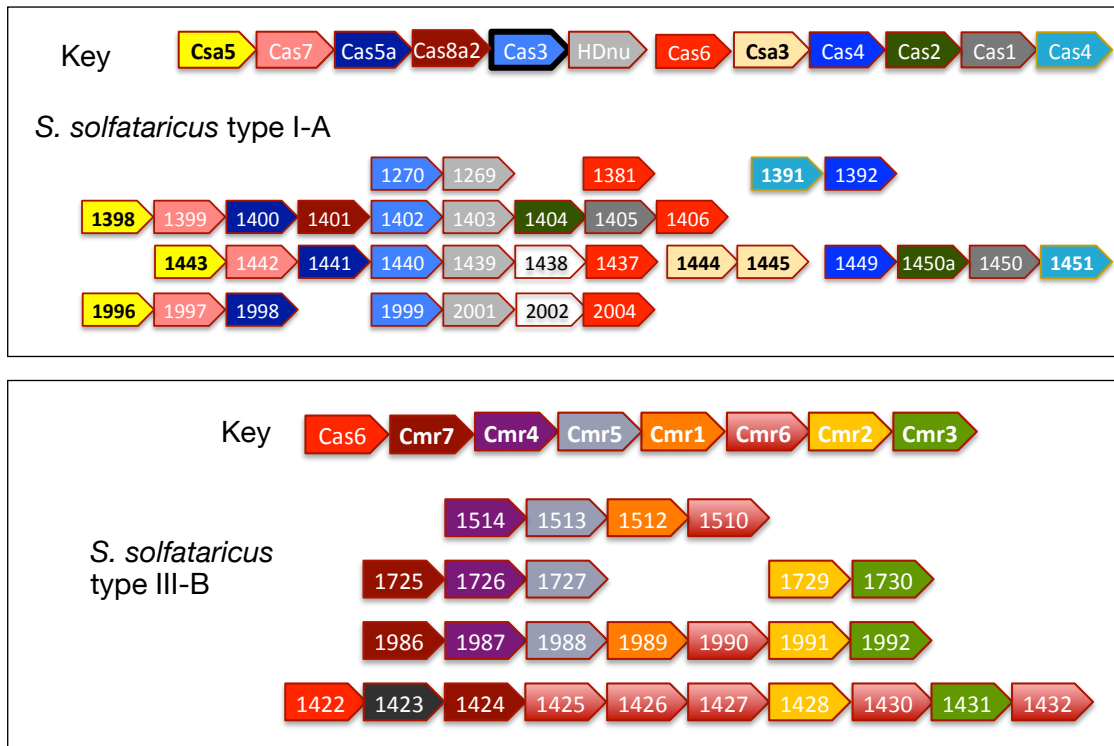


Figure 1.21: Cas genes in *S. solfataricus* P2

Operon organisation and gene names of the type I-A and III-B systems. Homologous genes are color-coded.

CRISPR locus	Family	repeats	genome location	consensus repeat
A (3)	II	103	1233466-1239959	GATTAATCCCAAAGGAATTGAAAG
B (4)	II	95	1254482-1260452	GATTAATCCCAAAGGAATTGAAAG
C (5)	I	32	1297153-1299148	GATAATCTCTTATAGAATTGAAAG
D (6)	I	96	1305539-1311637	GATAATCTCTTATAGAATTGAAAG
E (9)	I	8	1744007-1744417	GATAATCTACTATAGAATTGAAAG
F (10+11)	I	89	1809772-1815557	GCTAATCTACTATAGAATTGAAAG

Table 1.2: CRISPR loci in *S. solfataricus* P2

CRISPR loci names and families as in Lillestol *et al.* (2009). Numbers in parenthesis refer to the CRISPR database loci numbering (Grissa *et al.* 2007).

Gene name	COG	putative or confirmed activity	Stage involved	notes
Cas1	1518	ss and dsDNA endonuclease, binding RNA and DNA	adaptation?	Structure solved (3GOD, 3LFX, 2YZS)
Cas2	1343	endoribonuclease	adaptation?	Structure solved (2IVY, 2I8E, 3EXC)
Cas3'	1203	putative DExH-box helicase	interference	
Cas3'' (HDnuc)	2254	HD nuclease, cleaving dsDNA and dsRNA	interference	Structure solved (3M5F)
Cas4	1468	predicted RecB-like nuclease	adaptation?	
Cas5a	1688		interference	subunit of CASCADE-like complex
Cas6	1583	endoribonuclease	CRISPR transcript processing	Structure solved (3I4H)
Cas7	1857		interference	subunit of CASCADE-like complex
Cas8a2	?			subunit of CASCADE-like complex?
Csa3	0640	putative transcriptional regulator		Structure solved (2WTE)
Csa5	?		interference	subunit of CASCADE-like complex
Cmr1, Cmr3, Cmr4, Cmr6	RAMP		interference	subunit of CMR complex
Cmr2 (Cas10)	1353	contains HD nuclease, putative polymerase domain, Zn-ribbon	interference	subunit of CMR complex
Cmr5	3337		interference	subunit of CMR complex. structure solved (2OEB and 2ZOP)

Table 1.3: Characteristics of the Cas proteins in *S. solfataricus* P2

COG numbers from Makarova *et al.* (2006). Activities refer to either experimentally confirmed or predicted activities, as described in text. For available structures, PDB identifiers are mentioned in parentheses.

Chapter 2

Materials and Methods

2.1 Cloning procedures

2.1.1 Cloning and vectors

The *Sulfolobus solfataricus* P2 type III-B *cas* genes encoding Cmr1 (gene name sso1989), Cmr2 (sso1991), Cmr3 (sso1992), Cmr4 (sso1987), Cmr5 (sso1988), Cmr6 (sso1990) and Cmr7 (sso1986) were amplified from *S. Solfataricus* P2 genomic DNA by PCR and cloned into vectors pEHISTEV and pDEST14 to allow for polyhistidine-tagged protein expression. The sequences of the primers used are available upon request. Genes sso1986, 1988, 1990 were cloned into the 5' *Nco*I / 3' *Bam*HI restriction sites of the pEHISTEV vector (Liu and Naismith, 2009), while the genes sso1987, 1989, 1992 were cloned into the pDEST14 vector of the Gateway cloning system (Invitrogen). The rationale for the alternate cloning of sequential genes into these two vectors was to enable co-expression studies of various gene combinations by co-transformation of the two compatible vectors into the same host.

The Gateway cloning system is based on the site-specific recombination properties of bacteriophage lamda. A modified version of this system was developed and described by the SSPF (Oke et al. 2010), in which a third 5' common primer is used for the amplification of the gene of interest in addition to the two gene-specific primers. This 118bp common primer contains the attB1 recombination site and adds to the construct the following elements: a ribosome binding site, a transcription start codon, an N-terminal six-histidine tag and a TEV protease cleavage site.

The *cas* genes sso1440 (*cas*3'), sso1441 (*cas*5), sso1442 (*csa*2) and sso2004 (*cas*6) were also cloned into vector pDEST14 using the modified Gateway system. The sequence encoding for the first fourteen residues of the N-terminus of Sso1440 (Cas3') were removed, as comparative sequence analysis indicated that it is probably a misannotation event. To enable co-expression studies of sso1441 and sso1442, the two genes were cloned into a modified version of pRSFDuet-1 (Novagen), named pRSFDuetHISTEV (Oke et al. 2010), using restriction sites BspHI/BamHI and NdeI/

XhoI respectively. The modified vector enables the addition of a TEV cleavable six-histidine tag to the gene cloned into the first multiple cloning site. In this case Sso1441 would be expressed with the cleavable six-histidine tag while Sso1442 would be co-expressed in the native form. Cloning reactions were carried out by Dr. Huanting Liu (SSPF, University of St Andrews).

Cas genes *sso1437* (*cas6*) and *sso1443* (*csa5*) were cloned into expression vector pET151/D-TOPO using the TOPO TA Cloning system (Invitrogen) by honours student Maryam Qurashi.

All genes mentioned in this study were amplified from *Sulfolobus solfataricus* genomic DNA, isolated from exponentially growing cells using the animal tissue protocol of the QIAGEN DNeasy Tissue Kit. All primer sequences are available upon request and were purchased by Eurofins MWG Operon, Dharmacon and Eurogentec.

All plasmid constructs were sequenced at The Sequencing Service, School of Life Sciences, Dundee.

2.1.2 Site-directed mutagenesis

Site-specific mutants of the *sso1440* and *sso1442* genes were generated using the Stratagene QuikChange XL Site-Directed Mutagenesis Kit as per the manufacturer's instructions. The primer sequences used are available upon request. The base mutations were confirmed by DNA sequencing and the amino acid substitutions were verified by mass spectrometry of the purified protein.

The mutations were designed to target the conserved lysine in the Walker A motif of helicase Cas3' (*Sso1440*), and substitute it with an alanine, thus generating the mutant *Sso1440-K46A*. The mutations in *Csa2* (*Sso1442*) were designed to replace a conserved histidine at position 160 identified by multiple sequence alignment with an alanine residue, in an attempt to gain more information about the protein activity.

2.2 Protein expression and purification

2.2.1 Expression of recombinant proteins

The *Escherichia coli* competent cells BL21(DE3) and C43 (DE3) (Stratagene) were used for protein expression in all cases. The latter strain contains at least one uncharacterized mutation conferring tolerance to toxic proteins, and is derived from strain C41 (DE3), which in turn derives from BL21(DE3).

Luria-Bertani (LB; 10 g/l tryptone, 5 g/l yeast extract, 10 g/l NaCl) and Tryptone-phosphate broth (TPB; 20 g/l tryptone, 2 g/l K₂HPO₄, 2 g/l KH₂PO₄, 5 g/l

NaCl) were used as growth media supplemented with the appropriate antibiotic, either ampicillin or kanamycin at a final concentration of 100 µg/ml or 35 µg/ml respectively.

For large scale protein expression 10 ml overnight starter cultures were used to inoculate 1 litre LB or TPB (plus antibiotic) cultures in 2-litre flasks, which were incubated at 37°C, 180 rpm until the cells reached mid-log growth phase (OD₆₀₀ ~0.6-0.8). Protein expression was then induced with 0.4 mM IPTG and cultures were incubated overnight (~14 hours) at 25°C with shaking at 180 rpm. The cells were harvested by centrifugation at 6 krpm, 4°C for 15 min using a Beckman JLA 8.10000 rotor in an Avanti J-20 XP centrifuge and stored at -80°C until required. The expression conditions of the CAS proteins in this project are summarized in table 2.1. The cloned genes not mentioned in table 2.1 were also subjected to expression trials but were either insoluble or not expressed at all.

ORF (sso)	Name	Vector	Antibiotic resistance	Medium	Induction	post-induction growth conditions
1440	Cas3 ^{''}	pDEST14	Amp ^r	TPB	0.4mM IPTG	25°C o/n
1442	Csa2	pDEST14	Amp ^r	LB	0.4mM IPTG	25°C o/n
1441/1442	Cas5/ Csa2	pRSFDuetHI STEV	Kan ^r	LB	0.4mM IPTG	25°C o/n
1443	Csa5	pET151/D- TOPO	Amp ^r	LB	0.4mM IPTG	25°C o/n
1437	Cas6	pET151/D- TOPO	Amp ^r	LB	0.4mM IPTG	25°C o/n
1986	Cmr7	pEHISTEV	Kan ^r	LB	-	25°C o/n
1987	Cmr4	pDEST14	Amp ^r	LB	0.1 mM IPTG	25°C o/n
1989	Cmr1	pDEST14	Amp ^r	TPB	0.4mM IPTG	25°C o/n

Table 2.1 Expression conditions of Cas proteins from *S. solfataricus*

2.2.2 Purification of recombinant proteins

The general purification scheme employed in this study is the following, with the various adjustments described for each protein in the end. Cell pellets were resuspended at a 1:5 (w/v) ratio in affinity binding buffer (20 mM NaH₂PO₄/Na₂HPO₄ pH 7.4, 500 mM NaCl, 10 mM imidazole, 10% glycerol) supplemented with 1 mg/ml

lysozyme (Sigma), 50 µg/ml DNase I (Sigma) and protease inhibitor cocktail tablets ("Complete mini Protease Inhibitor mix EDTA-free" tablets, Roche Diagnostics, 1 tablet per 50 ml) prior to cell lysis by sonication for 12 min (6x2 min) on ice. The cell lysate was centrifuged for 30 min at 20 krpm, 4°C (Beckman JA 25.50 rotor) and the supernatant was filtered with a sterile syringe-driven 0.45 µm filter (Milipore) prior to purification by nickel-chelate affinity chromatography. The filtered lysate was loaded onto a 5 ml HisTrap HP Ni-sepharose column (GE Healthcare) pre-equilibrated with affinity binding buffer and the target protein was eluted over a 30 mM - 500 mM linear imidazole gradient. The protein-containing fractions were pooled together and subjected to cleavage of the six-histidine tag by the tobacco etch virus (TEV) protease in buffer containing 20 mM NaH₂PO₄/Na₂HPO₄ pH 7.4, 500 mM NaCl, 10% glycerol, 0.5 mM EDTA, 1 mM DTT, overnight at room temperature or 4°C, depending on the protein stability. The protein sample was then applied to an equilibrated with affinity binding buffer 5 ml HisTrap HP Ni-sepharose column (GE Healthcare) in order to remove the cleaved six-histidine tag, the histidine tagged TEV protease and other contaminants. The column flow-through, containing the untagged target protein, was concentrated to an appropriate volume and then loaded onto a HiLoad 26/60 Superdex 200 gel filtration column (GE Healthcare) equilibrated with the appropriate gel filtration buffer (depending on the protein and the subsequent purification step) from which the protein eluted as a single monodispersed peak. The gel filtration buffers used consisted of 20 mM MES pH 6 or 20 mM Tris-HCl pH 7.5, 500 mM NaCl, 1 mM EDTA, 0.5 mM DTT, 10% glycerol. All chromatographic purification steps were performed on a BioLogic DuoFlow chromatography system (Bio-Rad) and an AKTA Xpress automatic protein purification system. The efficiency of the procedure and the purity of the sample at each purification step was confirmed by SDS-PAGE electrophoresis on 4-12% NuPage gels (Invitrogen). The identity of the target protein was confirmed by MALDI-TOF and ESI mass spectrometry. Purified protein samples were pooled together, concentrated to an appropriate concentration with Vivaspin concentrators (Sartorius Stedim Biotech GmbH), flash-freezed in liquid nitrogen and stored at -80°C. The protein-containing samples were kept on ice at all times, unless otherwise stated.

Cas3' (Sso1440) (both wild-type and mutant versions) was purified as described above and stored in a final buffer containing 20 mM MES pH 6, 500 mM NaCl, 1 mM EDTA, 0.5 mM DTT, 10% glycerol. All purification procedures were carried out on ice due to partial protein degradation, unless otherwise stated.

Csa2 (Sso1442) was purified as described above with the addition of a third purification step, a 5 ml HiTrap Heparin HP column pre-equilibrated with a 20 mM MES pH 6, 50 mM NaCl, 1 mM EDTA, 0.5 mM DTT, 10% glycerol buffer, from which the protein eluted with a 50 mM - 1 M NaCl linear gradient. The six-histidine tag was uncleavable.

For the purification of the Csa2/Cas5 protein complex all procedures were carried out at room temperature and in the absence of DNase to prevent carry-over contamination. The soluble fraction was incubated at 65°C for 20 min to precipitate most of the *E. coli* contaminant proteins and centrifuged at 40 krpm in a Beckman Coulter Optima L-90K Ultracentrifuge with a Beckman 70Ti rotor prior to metal-chelate affinity purification. The six-histidine tag was not cleaved and a third purification step on a 5 ml HiTrap Heparin HP column was employed as described for Csa2. In the case of the Csa2/Cas5 complex the purification through a heparin column enabled also the separation of the biologically relevant complex dimer from the excess of expressed Csa2 subunits.

Cas6 (Sso2004) was purified by Dr Shirley Graham as described above and stored in 50 mM Tris-HCl pH 7.5, 150 mM NaCl, 10 % glycerol.

Cmr4 (Sso1987) and Cmr1 (Sso1989) were purified as described above, with the addition of 0.5% Triton X-100 in the lysis buffer.

Protein concentrations were calculated from the sample absorbance at 280 nM, measured on a Varian Cary 50Bio UV-Visible spectrophotometer, and the theoretical molar extinction coefficient obtained from the ExPASy ProtParam analysis of the protein sequence, using Beer Lambert's law. In the case of Csa2 (Sso1442) which lacks tryptophan residues, the Bradford reagent (Bio-Rad) was also used to calculate protein concentrations as per the manufacturer's instructions using a BSA standard curve.

2.3 Crystallization screening and optimisation

The optimal concentration for crystallisation screens was determined by performing the Pre-crystallization Test (Hampton Research), which evaluates the sensitivity of the protein to salt and polymer concentrations and the homogeneity and purity of the sample.

Commercial sparse matrix crystallisation screens (Classics and JCSG by Qiagen, JMac, Wizard by Emerald Biosystems, PEGs and JMac by Hampton Research, Proplex by Molecular Dimensions and stochastic screens prepared by the SSPF) were used to determine initial crystallisation conditions. The sitting drop vapour-diffusion method was used for all the initial crystallisation experiments. Screening experiments with 300 nl drops (150 nl protein plus 150 nl crystallisation buffer) or 450 nl drops (300 nl protein plus 150 nl crystallisation buffer) were set by the Cartesian Honeybee nanodispenser robotic system (Genomic Solutions), or 2 µl drops (1 µl of protein solution plus 1 µl of crystallisation buffer) were set manually on 96-well crystallisation plates. The crystal trays were sealed, kept at 20°C and examined regularly under an optical microscope or using the Rhombix Vision visualisation robot.

Conditions resulting in the formation of some form of microcrystals or crystalline precipitant were selected for optimisation trials, in which a grid screen or a stochastic screen was designed around the initial conditions in order to optimise crystal formation. The screens were set manually using the hanging-drop vapour diffusion method in 24-well trays. The drops were formed by mixing 2 μ l and 1 μ l of protein solution with 1 μ l and 0.5 μ l of crystallisation buffer respectively, and the reservoir volume was set to 450 μ l. The crystal trays were sealed, kept at 20°C and examined under an optical microscope.

All the screening experiments were carried out in collaboration with the SSPF. The SSPF personnel also handled the crystal harvesting, data collection and structure solution of the protein targets.

2.4 Immuno - blot

In order to identify fractions containing the Cmr protein complex in *S. solfataricus* extract through sequential purification steps, the Dot blot method was employed using antibodies raised against the desired component protein.

Sheep antiserum containing polyclonal antibodies against Cmr7 was purchased by the Scottish National Blood Transfusion Service. Two microlitres of each fraction were blotted on a HyBond ECL nitrocellulose membrane (Amersham Biosciences) and incubated in blocking buffer consisting of PBS, 5% milk powder, 0.5% Tween-20 (Sigma) for 10 min. To enable the antibody-protein interaction the membrane was incubated in a 1:2000 or 1:1000 dilution of the primary antibody (anti-Cmr7) in PBS, 2% milk powder, 0.2% Tween-20 for 30 min to 1 hour. After washing 3x10 min with fresh blocking buffer the membrane was incubated in a 1:10,000 dilution of the secondary antibody (rabbit anti-goat, ImmunoPure Antibody, Pierce) for 30 min. Finally the membrane was washed in blocking buffer (3 x 10 min) and ultrapure water (3 x 5 min). All the incubation and washing steps were performed at room temperature with gentle shaking. For the detection of the chemiluminescent signal the SuperSignal West Pico Chemiluminescent Substrate or Femto Maximum Sensitivity Substrate Kits (Pierce) were used as per the manufacturer's instructions. The signal was visualized on a Fuji Luminescent Image Analyser LAS-1000 and Image Reader software.

2.5 Protein Interactions

2.5.1 Analytical size exclusion chromatography

Gel filtration chromatography can be used for the detection and characterization of stable protein-protein interactions. The formation of a stable

protein complex would result in different partitioning between the mobile and the stationary phase of the column compared to the complex components, due to the larger Stokes' radius of the complex.

A Superose™ 12 10/300 GL (Tricorn™) (GE Healthcare) was equilibrated with 20 mM MES pH 6, 250 mM NaCl, 10% glycerol, 1 mM EDTA, 0.5 mM DTT and calibrated with known molecular weight standards (blue dextran, β -amylase, alcohol dehydrogenase, albumin, carbonic anhydrase and cytochrome C (Sigma)) in order to obtain a linear calibration curve of the logarithms of the known molecular weight standards against their respective V_e/V_o values. The protein samples (either single purified protein samples in order to determine their oligomeric state or mixed samples of potentially interacting proteins) were run through the column at a flow rate of 0.5 ml/min and the eluted peak fractions were analysed by SDS PAGE and mass spectrometry. The apparent molecular weights of the loaded samples were calculated using their elution volumes (V_e) and the calibration curve.

2.5.2 Determination of protein interactions using magnetic precharged nickel particles

Interactions between recombinant proteins can also be identified by making use of the six-histidine tag, in the same principle as nickel-chelating affinity chromatography. The histidine tagged protein, bound to the nickel-loaded particles, can be used as “bait” to affinity purify any interacting proteins from a pool of proteins in solution. To identify interactions between the recombinant Cmr subunits, Ni-NTA magnetic agarose beads (Qiagen) were used. In an eppendorf tube, 25 μ l of magnetic agarose beads with covalently bound nickel chelating nitrilotriacetic acid (NTA) groups were incubated with 7.5 μ g of polyhistidine-tagged protein in 25 μ l of affinity binding buffer (10 mM Tris-Cl pH 8, 100 mM NaCl, 10 mM imidazole, 5 mM $MgCl_2$), for 1 hour at room temperature. After removal of the unbound protein, the interacting partner was added to the beads at an equimolar concentration and incubated for 1 hour at room temperature. After a series of washes with increasing NaCl (250 mM and 500 mM) and imidazole concentration (50 mM and 100 mM), the tagged protein is eluted with 25 μ l of 10 mM Tris-Cl pH 8, 500 mM NaCl, 500 mM imidazole, 5 mM $MgCl_2$. Any remaining protein is eluted further by incubating the beads at 95°C for 5 min and the samples are analysed by SDS-PAGE. Non-specific interactions between the non-tagged proteins and the Ni-NTA beads were also tested. If the tagged and untagged proteins are forming a complex under the assay conditions, they would be found in the elution fraction.

To characterize interactions between recombinant Csa2, Cas5, Cas6, Csa5, Cas3' in the presence or absence of crRNA, MagneHis Ni-Particles (Promega) were used. These experiments were carried out in collaboration with Dr Shirley Graham. In

an eppendorf tube, 20 μ l of magnetic nickel-loaded particles were incubated with 12 μ g of polyhistidine-tagged protein in affinity binding buffer for 10 min at 40°C. The unbound protein was removed and the bound “bait” protein was incubated with an equimolar concentration of the potential interacting partner(s) in the presence or absence of crRNA for 10 min at 40°C. In the samples containing Cas3', ATP / MgCl₂ was also added at a concentration of 1 mM. The magnetic beads were then washed with increasing concentrations of imidazole and NaCl and finally the bound proteins were eluted from the beads with 25 μ l of 10 mM Tris-Cl pH 8, 500 mM NaCl, 500 mM imidazole. The samples were analysed by SDS-PAGE. Where indicated, the beads were resuspended in SDS-PAGE loading buffer and incubated at 90°C for 5 min before loading on the gel. This final step ensured that all the tightly bound proteins would be removed from the beads.

2.6 Generation of nucleic acid substrates and markers

2.6.1 Purification of oligonucleotides

Synthetic RNA and DNA oligonucleotides (purchased from Integrated DNA Technologies and Eurofins MWG Operon respectively) were denatured in 50% formamide (Promega) at 80°C for 10 min and transferred on ice until separated on a denaturing 12% polyacrylamide / 7 M urea gel for 90 min at 22W. The gel was pre-run for 30 min and wells were thoroughly rinsed before samples were loaded. Bands were visualised by UV shadowing, excised from the gel and incubated overnight at 4°C in TE buffer (10 mM Tris pH 7.5, 1 mM EDTA). The DNA or RNA was purified by ethanol precipitation, resuspended in TE buffer or the RNA Storage Solution (Ambion) for RNA oligonucleotides, and stored at -20°C.

Name	Sequence 5' to 3'
CRISPR_compF DNA	CTTTCAATTCTATAGTAGATTAGC
CRISPR_compF RNA	CUUUCAAUUCUAUAGUAGAUUAGC
CRISPR_compB DNA	CTTTCAATTCCTTTTGGGATTAATC
CRISPR_compB RNA	CUUUCAAUCCUUUUGGGAUUAAUC

Table 2.2: Oligonucleotides used in chapter 3

Name	Sequence 5' to 3'
crRNA-A1	AUUGAAAGGAACUAGCUUAUAGUUUAGAAGAAAACAAACAAAUA AU GAUUAUCCCAAAA
crRNA-A1_Δ3	AUUGAAAGGAACUAGCUUAUAGUUUAGAAGAAAACAAACAAAUA AU
crRNA-A1_Δ5	GAAAGGAACUAGCUUAUAGUUUAGAAGAAAACAAACAAAUAU GAUUAUCCCAAAAAGGA
A1P control	AGGGUAUUUUUGUUUUUCUUCUAAACUUAAGCUAGUUC
tA1f +PAM	TAATACGACTCACTATAGGGT ATTATTTGTTTGTTCCTTCTAACTATAAGCTAGTTCTGGAGA GAAGGTG
tA1r +PAM	CACCTTCTCTCCAGAACTAGCTTATAGTTTAGAAGAAAACAAAC AAATAATACCCTATAGTGAGTCGTATTA
tA1f-PAM	TAATACGACTCACTATAGGGTATTATTTGTTTGTTCCTTCTAA ACTATAAGCTAGTTCCCAAGAGAGGTG
RNA_tA1f	AGGGUAUUUUUGUUUUUCUUCUAAACUUAAGCUAGUUCU GGAGA
crRNA native	AUUGAAAGGAACUAGCUUAUAGUUUAGAAGAAAACAAACAAAUA AUGAUUAUCCCAAAAAGGA
U15 CRISPR repeat locus B	UUUUUUUUUUUUUUUGAUUAUCCCAAAAAGGAAUUGAAAG

Table 2.3: Oligonucleotides used in chapter 4

Name	Sequence 5' to 3'
CRISPR locus B DNA	CTTTCAATTCCCTTTTGGGATTAATC
CRISPR locus B RNA	CUUUCAAUCCUUUUUGGGAUUAUC
CRISPR 3' oh DNA	GATTAATCCCAAAGGAATTGAAAGTTTTTTTTTTTTTTTT
CRISPR 5' oh DNA	TTTTTTTTTTTTTTTTTGGATTAATCCCAAAGGAATTGAAAG
CRISPR 5' oh RNA	UUUUUUUUUUUUUUUGAUUAAUCCCAAAGGAAUUGAAAG
CRISPR B complement	GATTAATCCCAAAGGAATTGAAAG
MKDNAf	GCTCCTAGGTCCTTCGTGGCATCTG
MKDNAr	CGAGGATCCAGGAAGCACCGTAGAC
MKDNA5'oh	AAAACAAAACAAAATCAGATGCCACGAAGGACCTAGGAGC
MKDNA3'oh	CAGATGCCACGAAGGACCTAGGAGCTAAAACAAAACAAAA
MKRNAf	GCUCCUAGGUCCUUCGUGGCAUCUG
MKRNAr	CGAGGAUCCAGGAAGCACCGUAGAC
MKRNA5'oh	AAAACAAAACAAAUCAGAUGCCACGAAGGACCUAGGAGC
MKRNA3'oh	CAGAUGCCACGAAGGACCUAGGAGCUAAAACAAAACAAAA
MKDNAcomp	CAGATGCCACGAAGGACCTAGGAGC
MKRNAcomp	CAGAUGCCACGAAGGACCUAGGAGC

Table 2.4: Oligonucleotides used in chapter 5

2.6.2 Assembly and purification of double-strand substrates

The purified DNA or RNA oligonucleotides designed to serve as substrates for catalytic assays were either already purchased with a 5' fluorescein label or 5' end labelled with [γ - 32 P] ATP (4500 Ci / mmol, MP Biomedicals) using T4 PNK (Fermentas) as per the manufacturer's instructions. The labelled oligonucleotides were annealed into various structures (Appendix I) by slow cooling from 85°C to room temperature overnight and purified with 10% Ficoll on a native 12% polyacrylamide / 50 mM NaCl gel. The duplex constructs were eluted overnight at 4°C in TE / 50 mM NaCl, ethanol precipitated, resuspended in TE - 50 mM NaCl and stored at -20°C.

2.6.3 CRISPR locus constructs

Part of the *S. solfataricus* CRISPR locus A consisting of the leader sequence directly upstream of the repeats and 4 repeat-spacer units was amplified from *S. solfataricus* genomic DNA by polymerase chain reaction. Two constructs were generated corresponding to either the whole leader sequence (245bp upstream from the start of the CRISPR locus) or part of it (165bp upstream) and four repeat-spacer units, named CRISPR I, II (figure 2.1) The amplified CRISPR fragments were cloned into pCR2.1 TOPO vector as per the manufacturer's instructions, transformed into competent *E. coli* DH5a cells and isolated plasmids from the resulting colonies were sequenced to verify the presence and orientation of the insert.

A different construct containing the T7 promoter sequence directly upstream of the first two repeat-spacer units of CRISPR locus A was generated by PCR in order to serve as template for *in vitro* transcription (CRISPR T7). The generated sequences can be found in Appendix I.

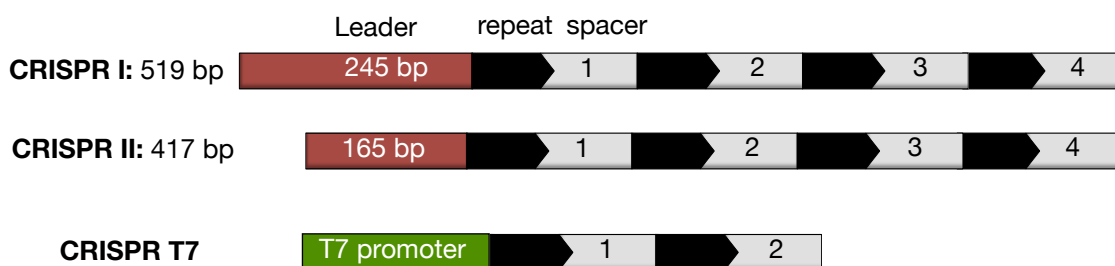


Figure 2.1: Amplified fragments of the *S. solfataricus* CRISPR locus A

Leader sequence is illustrated as a red box, repeat sequences as black arrows and spacer sequences as numbered light gray boxes. The T7 promoter is illustrated as a green box. Relative sizes are not up to scale.

2.6.4 T7 RNA polymerase-mediated *in vitro* transcription

The T7 CRISPR PCR fragment was purified by gel extraction (Qiagen Gel Extraction Kit) and used as template for *in vitro* transcription using T7 RNA polymerase Plus (Ambion) as per the manufacturer's instructions. The crRNA transcript was uniformly labelled with [α - 32 P] UTP (3000 Ci / mmol, MP Biomedicals) and purified on a 15% denaturing polyacrylamide/urea gel. Before loading the sample was heated at 65°C for 15 min to denature any secondary structures and mixed with formamide loading dye. The transcript was ethanol precipitated and resuspended in TE buffer or the RNA storage solution (Ambion). The transcript concentration was too low to measure by the sample absorbance at 260 nM.

2.6.5 Sanger DNA sequencing

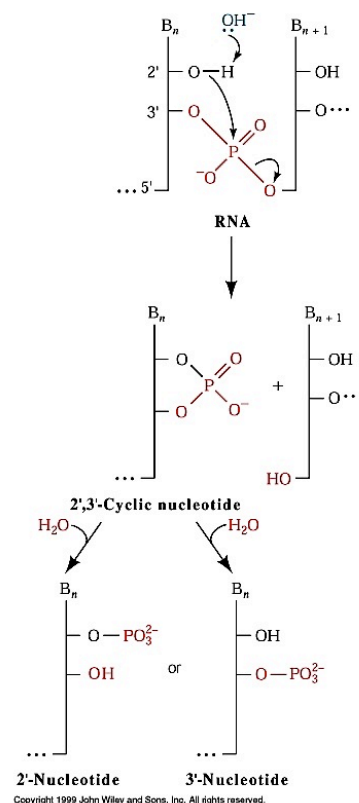
The CycleReader™ DNA Sequencing Kit (Fermentas) was used to sequence the amplified CRISPR locus fragments and map the transcription initiation site within the CRISPR leader sequence. This method is based on Sanger sequencing of the DNA template, where the polymerase-mediated synthesis of a new strand begins at the 3' end of a primer annealed to the template DNA and is terminated by incorporation of dideoxynucleotides (ddNTPs), which cannot form subsequent phosphodiester 3' - 5' bonds as they lack the 3' hydroxyl group. By performing four separate reactions with the four different ddNTPs, a population of DNA strands complementary to the template strand is generated with a fixed 5' end and variable 3' ends, covering the whole length of the template sequence. By comparing the length of the terminated strands with the original transcript produced by the same template, the native transcription initiation site can be determined.

20 - 60 fmol of the CRISPR locus construct II described in paragraph 2.6.3 was used as sequencing template (either as PCR fragment or linearized plasmid construct) and 20 pmoles of the reverse complement oligonucleotide for spacer 1 was used as a primer for the kit's thermostable *Taq* DNA polymerase. The thermal sequencing reactions were carried out in a Techne TC-512 thermal cycler as described in the manufacturer's manual using the direct label incorporation protocol, which enables the incorporation of [α - 32 P] dATP in the produced strands. Reaction products were analysed on a denaturing 15% PAA, 7 M urea gel.

2.6.6 RNA alkaline hydrolysis ladder

To generate a gel electrophoresis ladder of hydrolyzed RNA fragments 0.1-3 μ g of end-labelled RNA were incubated in 50 mM [NaHCO₃ / Na₂CO₃] pH 9.2, 1 mM EDTA in a final volume of 10 μ l, for 15 min at 95°C. Samples were mixed with 2x formamide-based gel loading dye (95% formamide, 0.025% bromophenol blue, 0.025% xylene cyanol FF, 5 mM EDTA pH 8, 0.025% SDS) and stored at -20°C. The RNA fragments generated with this reaction have the phosphate group attached to the 3' hydroxyl group (figure 2.2).

Figure 2.2: Mechanism of RNA hydrolysis under alkaline conditions



2.7 *Sulfolobus solfataricus* *in vitro* transcription

To determine whether the leader sequence of CRISPR locus A contains a canonical promoter sequence that could be responsible for transcription *in vivo*, we performed *in vitro* transcription using the *S. solfataricus* system components. The pCR2.1 TOPO vectors containing the CRISPR locus constructs described in 2.6.3 were digested with *Hind*III (Fermentas) to produce linear DNA template for transcription. The reaction mix consisting of 100 ng linearised plasmid template, 2x transcription buffer (40 mM Tris-HCl pH 8, 20 mM MgCl₂, 4mM DTT, 440 mM KCl), 40 nM SsoRNAPol, 80 nM SsoTFB-1, 80 nM SsoTBP, 300 nM BSA was incubated at 70°C for 10 min to allow for the transcription initiation complex to form. The reaction was initiated with the addition of rNTPs (200 µM each) in a final reaction volume of 50 µl and incubated for a further 20 min. The *S. solfataricus* RNA polymerase, TFB-1 and TBP were kindly provided by Dr Sonia Paytubi.

The transcribed RNA was subsequently used as a template for a primer extension reaction. The primers used were complementary to the final spacer sequence of each cloned CRISPR fragment (CRISPR +252r, CRISPR+126r, see Appendix I) and were 5' end labelled with [γ -³²P] ATP. An appropriate volume of the transcription reaction and 300 fmol of the labelled reverse primer were incubated at 70°C for 5 min to enable hybridisation with the RNA transcript and chilled on ice. The primer extension reaction mix (5x Fermentas RevertAid H-Minus MMuLV RT buffer, 25 mM dNTPs, 4u RNasin (Promega)) was added to the hybridised transcript and primer solution and incubated for 5 min at 37°C. 1 µl of RevertAid H-Minus MMuLV Reverse Transcriptase (Fermentas, 200 u/µl) were added and the reactions were incubated at 42°C for 1 hour. The products were separated on a 20 % polyacrylamide / 7M urea gel, ran at 92W, 45°C for 90 min. The gel was exposed to a phosphoscreen overnight and then visualised by phosphorimaging on a Fujifilm FLA500 scanner using ImageGauge software (Fuji).

2.8 Extraction of RNA from purified native aCASCADE

Samples of aCASCADE purified from *Sulfolobus solfataricus* P2 cells were kindly provided by Dr Nathanael Lintner and Dr Martin Lawrence in Montana State University Bozeman. 20 µl samples of purified aCASCADE at 1-3 mg/ml concentrations were diluted with DEPC-treated water to 100 µl volume, to which 1 µl RNAsin and 100 µl phenol (Sigma) were added. After mixing the samples for 1 min by vortexing and centrifuging for 1 min at room temperature, the aqueous phase was transferred to a new eppendorf tube with 100 µl 24:1 (v/v) chloroform:isoamylalcohol (Sigma). The last step was repeated, and the aqueous phase was again transferred to a new eppendorf tube. Nucleic acids were precipitated for 1 hour at 80°C with 12 µl of

3 M sodium acetate and 250 μ l 100% ethanol, centrifuged for 30 min at 4°C and the pellet washed with 75 % ethanol. Centrifugation was repeated and the pellet was air-dried and resuspended overnight at 4°C in 10 μ l DEPC-treated water. The extracted nucleic acids were labeled with [γ -³²P] ATP as described in previous paragraphs and analysed on denaturing 20 % polyacrylamide, 7M urea gel, ran at 92W, 45°C for 90 min. Gel was exposed to a phosphoscreen overnight and then visualised by phosphorimaging on a Fujifilm FLA500 scanner using ImageGauge software (Fuji).

2.9 Nucleic acid binding and catalytic assays

2.9.1 Helicase assays

Helicase activity assays for Cas3' were carried out in 20 mM MES pH 6.5, 100 mM potassium glutamate, 1 mM MgCl₂, 1 mM ATP, 0.1 mg/ml BSA, 25 nM fluorescein or ³²P-labelled duplex DNA or RNA substrate and over a protein concentration range of 20 nM to 250 nM. The reaction mix (minus ATP / MgCl₂) was incubated at 37°C (unless otherwise stated) for 1 min and the reaction started with the addition of an ATP / MgCl₂ mix to a final concentration of 1 mM. 10 μ l samples were removed at appropriate time points and immediately added to 20 μ l of chilled stop solution consisting of 10 mM Tris-HCl pH 8, 5 mM EDTA, 0.5% SDS, 1mg/ml Proteinase K, 5 μ M competitor DNA (unlabelled DNA complementary to the displaced strand to prevent reannealing). Samples were incubated at room temperature for 15 min to allow proteinase K digestion, mixed with 10% Ficoll and loaded onto a native 12% polyacrylamide:TBE gel. The gel was run at 150V for 2-3 hours and visualised by phosphorimaging on a Fujifilm FLA500 scanner. In the case of ³²P-labelled substrates, the gels were first exposed to phosphoscreens overnight and then visualised by phosphorimaging. ImageGauge software (Fuji) was used to quantify the data and estimate the percentage of duplex unwinding as the ratio of (unwound/total substrate) x100. KaleidaGraph software was used to plot the unwound (%) substrate over time.

The following control reactions were prepared: i) without protein and ATP and incubated at the assay temperature to indicate substrate stability, ii) without protein and ATP and incubated at 95°C for 2min prior to cooling on ice to indicate substrate melting, iii) without ATP to show ATP activity dependence of the protein.

2.9.2 Endonuclease assays

Nuclease activity of Cas6 (and potential activity of Csa2) was assayed in reaction buffer (20 mM MES pH 6.0, 100 mM potassium glutamate, 0.5 mM DTT, 5 mM EDTA) at 45°C or 60°C for 15-30 min with 1 μ M recombinant protein and 100 nM synthetic RNA substrate or 1 μ l crRNA transcript (unknown concentration). At

appropriate time points 10 μ l aliquots were removed and treated with 0.1 mg Proteinase K for 15 min at room temperature. The samples were mixed with 10 μ l formamide loading dye, heated at 65°C for 5 min to denature secondary structures and the products were separated on 20 % polyacrylamide, 7M urea gels, ran at 92W, 45°C for 90 min. Cleavage products visualized by phosphorimaging, as described in the previous paragraph and quantified with ImageGauge software.

2.9.3 ATP hydrolysis reaction

The Malachite Green Phosphate Assay Kit (BioAssay Systems) was used to characterize the ATP hydrolysis activity of Cas3'. A standard curve of free phosphate pmoles over OD₆₂₀ was made with serial dilutions of K₂PO₃ as per the manufacturer's instructions. Reactions were performed in 20 mM MES pH 6.5, 100 mM potassium glutamate, 1 mM MgCl₂, 1 mM ATP, 0.1mg/ml BSA, 2 μ M protein and 2 μ M nucleic acid substrates (unless otherwise stated). Reactions were incubated at 55°C for 1 min and initiated by the addition of 1 mM ATP / MgCl₂. At appropriate time points 40 μ l samples were removed and immediately added to 40 μ l chilled 0.3M perchloric acid in a 96 well plate. Once all samples were collected, 20 μ l malachite green solution were added to each sample and incubated at room temperature for 12 min. Absorbance at 620 nm was measured on a SpectraMAX 250 Microplate Reader (Molecular Devices). All reactions were performed in triplicate. Appropriate control reactions without protein were performed and subtracted as background. The rate of ATP hydrolysis was plotted as pmoles of free phosphate over time with KaleidaGraph software.

2.9.4 Electrophoretic Mobility Shift Assay (EMSA)

To detect DNA or RNA binding by the CAS proteins gel mobility shift assays were performed, in which protein-nucleic acid complexes are observed as their rate of migration on native polyacrylamide gels is slower than for free oligonucleotides.

In the case of the Csa2-Cas5 complex, radiolabelled substrate (25 nM final concentration) was pre-incubated in binding buffer (20 mM MES pH 6.0, 50 mM potassium glutamate, 0.5 mM DTT, 5 mM EDTA, 5% glycerol and 100 nM unlabeled crRNA where indicated) at 55°C for 1 min and reactions were initiated by the addition of the Csa2/Cas5 complex at different concentrations between 250 nM - 5 μ M. Where unlabeled crRNA was present, the protein was pre-incubated with the unlabeled RNA prior to the addition of the labeled substrate. Reactions were incubated at 55°C for 10 min, mixed with 10% ficoll loading dye and separated on native 10% polyacrylamide gels at 120V for 3 hours. Results were visualised by phosphorimaging as described for helicase assays.

To characterize the DNA binding abilities of the recombinant CMR complex subunits, serial dilutions of the proteins were incubated in 50 mM Tris-HCl pH 8, 100 mM KCl, 5 mM MgCl₂, 1 mM DTT at 37°C or 50°C for 20 min with 1 μM labelled DNA or RNA substrate. Samples were mixed with 10% ficoll loading dye, separated on native 8% polyacrylamide gels at 120V for 2-3 hours and visualised by phosphorimaging.

2.9.5 Strand annealing and strand exchange assays

Strand annealing reactions for Cas3' were carried out in 20 mM MES pH 6.5, 100 mM potassium glutamate, 1 mM MgCl₂, 0.1 μl RNAsin (RNase inhibitor, Ambion Biosciences), 10 nM individually purified, 25 nt DNA or RNA strands and 1 μM protein. Only one of the strands was 5' fluorescein-labeled. The protein was incubated with the first strand at 37°C for 3 min, and the second strand was added to initiate the reaction. At indicated time points, 10 μl samples were removed and added to 15 μl chilled stop solution consisting of 10 mM Tris-HCl pH 8, 5 mM EDTA, 0.5% SDS, 1 mg/ml Proteinase K, 5 μM competitor DNA (unlabelled DNA complementary to the unlabelled strand to prevent reannealing). Samples were incubated at room temperature for 15 min to allow proteinase K digestion, mixed with 10% Ficoll and loaded onto a native 12% polyacrylamide:TBE gel. The gel was run at 150V for 2-3 hours and visualised by phosphorimaging on a Fujifilm FLA500 scanner.

For strand exchange reactions, 500 nM of protein were incubated with 50 nM unlabelled single strand substrate in 20 mM MES pH 6.5, 100 mM potassium glutamate, 1 mM MgCl₂, 50 mM NaCl, 0.1 μl RNAsin for 3 min at 37°C or 30°C as indicated. Reactions were initiated with the addition of 30 nM labelled double strand substrate and 1 mM ATP where indicated. At specific time points, 10 μl samples were removed, added to 15 μl chilled stop solution (10 mM Tris-HCl pH 8, 5 mM EDTA, 0.5% SDS, 1 mg/ml Proteinase K) and processed as in annealing reactions. Substrates were designed so that initial and product double strand species would be different in size.

2.9.6 R-loop formation assay

To investigate whether Cas3' would promote invasion of a ssRNA into a complementary dsDNA substrate to form an R-loop, the following reactions were carried out. 5 μM of Cas3' were incubated with 200 ng supercoiled pET151/D-TOPO plasmid containing a 25 bp protospacer region in 20 mM MES pH 6.5, 100 mM potassium glutamate, 0.1 μl RNAsin, 0.5 mM DTT at 55°C for 5 min. To initiate the reaction, 1 μM of 5' [³²P]-end labelled ssRNA containing a complementary spacer region was added to a final reaction volume of 10 μl. After 1 hr incubation at 55°C,

reactions were added to 15 μ l chilled stop solution (10 mM Tris-HCl pH 8, 5 mM EDTA, 0.5% SDS, 1 mg/ml Proteinase K) and incubated at 37°C for 15 min to allow for Proteinase K digestion. 2 μ l of loading dye (60 % glycerol, 0.25 % bromophenol blue, 0.25 % xylene cyanol) were added to the reactions and products were analysed by electrophoresis on a 0.8 % agarose-TBE gel at 90 V for 3 hrs . Gels were visualised by EtBr staining/ UV imaging and phosphorimaging. Reactions were also carried out in the presence of 5 μ M recombinant or native Csa2-Cas5a complex and 1 mM ATP / MgCl₂ (added at the reaction initiation stage) where indicated.

Chapter 3

The CMR complex from *Sulfolobus solfataricus*: native isolation and recombinant components

3.1 Introduction

3.1.1 The Repeat-Associated Mysterious Proteins (RAMPs)

A distinct set of CAS proteins belong to the superfamily of Repeat-Associated Mysterious Proteins (hereafter referred to as RAMPs), as identified by Makarova *et al.* in 2002. The discovery of this superfamily ensued initially as a result of the sequence analysis and secondary structure prediction of proteins from COGs 1336, 1367, 1604, 1337, 1332 that were associated with what was then thought to be a novel prokaryotic DNA repair system. Members of this superfamily show extremely weak sequence conservation but appear to share the same fold, a fact which was confirmed when the first CMR protein structures were solved (discussed later). Five specific structural motifs were identified by multiple alignment of members of this superfamily, including a β -strand followed by a conserved glycine near the N-terminus (motif I), a loop followed by an α -helix (motif II) and a C-terminal glycine-rich loop (motif V) (figure 3.1) (Makarova *et al.*, 2002). Subsequent analysis and identification of more RAMP protein families revealed that this is the most diverse and over-represented CRISPR-associated superfamily, with representatives found in all CAS subtypes.

The RAMP protein structures solved to date happen to be also the processing endonucleases of various CRISPR/CAS subtypes, which process the CRISPR RNA transcripts into the small crRNA effector sequences that perform the invader silencing. The first structure to be solved by x-ray crystallography was the CAS protein TTHB192 from *Thermus thermophilus* HB8 (CasE/Cse3/Cas6e homolog, YgcH-like family, Ecoli subtype, Ebihara *et al.* 2006).

MOTIFS	specific	I	II	III	IV	V
COGs 1336,1367, 1604,1337,1332		h.h...s.h.hg.s	ust-lkGhh+.hh	hhGtt	h.D	lGht.t.g.g.h
y1726-like			slhlpEKIVRGT		lRTIDg	YGUVTs.Ghuh
COG1851			hGchpG.psaFh			hGFRh
BH0337-like			h.pA-h+Gih-qih		hhLpDV	LGsREh.u.ht
COG1567		.hhhp.p	up.s-lhAh...h			lus.c.o.GhG.h
COG1769		hhh+h-.hhh	.s.s-hhGhls.h	h.G.h		hGtcp+sttchp
COG1688 (Cas5)		h..h...hh.ht.s	sa.sshhGhl...sh			lGttp..h.h
COGs 1583,5551	hh.hhPphl					hGtppshGFG.l
YgcH-like	hHphlh					hC.u+uhGhGhh
y1727-like	LHphLh					.G.FsthGLStss
MJ0978-like	hNH					lC+tsuhGhC.l

Figure 3.1: Conserved RAMP superfamily motifs and individual RAMP families

Adapted from Makarova *et al.* (2006). h: hydrophobic residue, p: polar residue, t: residue with high turn-forming propensity, +: positively charged residue.

The structure contained a double ferredoxin-like fold with the signature $\beta\alpha\beta\beta\alpha\beta$ secondary structure arrangement in each domain, and an RRM motif formed from the basic surface of the four-stranded β -sheet. The conserved Gly-rich loop was inserted between the last α -helix and β -strand of the polypeptide chain (figure 3.2). These domains are generally associated with RNA-binding proteins (e.g. ribosomal proteins S6 and S10, spliceosome subunits) (Maris *et al.* 2005). This fact, along with the results of various structure similarity search programs, led Makarova *et al.* (2006) to speculate that RAMPs function as RNA-binding proteins specializing in crRNA fragments of different sizes, and could form hetero-oligomeric complexes, analogous to the RNA-binding proteins of the eukaryotic RISCs implicated in the eukaryotic RNAi.

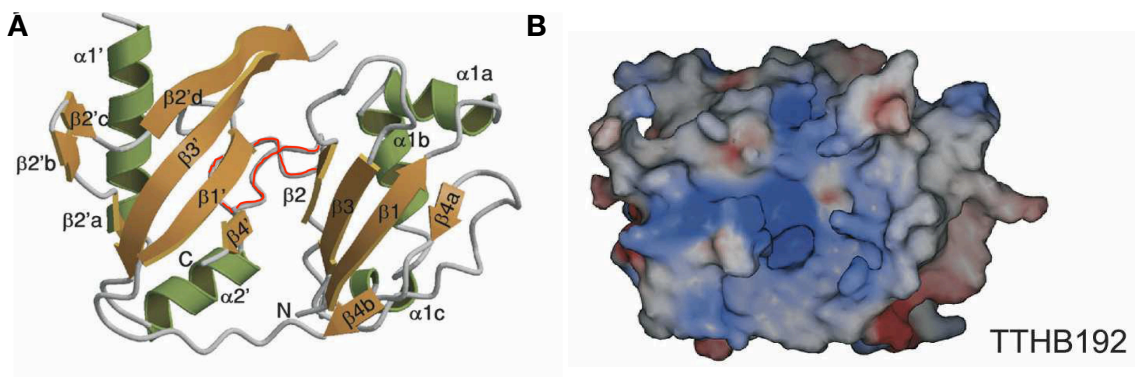


Figure 3.2: Crystal structure of Tthb192 - CasE/Cse3 homologue

(A) Ribbon diagram of the Tthb192 structure, adapted from Ebihara *et al.* (2008). α -helices are colored in green, β -strands in orange and numbered from the N to the C-terminus. The N and C termini are labeled, and the G-rich loop is highlighted in red. (B) Electrostatic surface potential illustrating the conserved basic patch on the β -sheet platform of the protein. Blue shaded areas indicate the positively charged regions ($+20k_B T$) and red areas the negatively charged regions ($-20k_B T$).

Subsequent crystal structures of RAMP proteins such as Cas6 from *Pyrococcus furiosus* (PfuCas6, Carte *et al.* 2008; Carte *et al.* 2010; Wang *et al.* 2011) and Csy4 from *Pseudomonas aeruginosa* (Haurwitz *et al.* 2010) revealed the ferredoxin-like folds and RRM motifs that seem to be characteristic of this superfamily, although their sequence identity is below 10%. While this seems to suggest a similar mechanism for RNA processing, biochemical and structural studies of these proteins along with co-crystals obtained with crRNA sequences from the respective organisms illustrated the distinct recognition and catalytic mechanisms these proteins employ *in vivo* (discussed in chapter 1). This is to be expected, considering the different CRISPR/CAS subtypes the proteins belong to and the different properties of the CRISPR repeats they encounter. Cas6 is associated with subtypes I-A, I-B, I-D, III-A, III-B while Csy4 with subtype I-F (Makarova *et al.* 2011a). All three proteins use distinct sequence and structure-specific recognition mechanisms to discriminate their respective substrates, illustrating the versatility of the characteristic duplicate ferredoxin-like fold in RAMPs and providing a mechanistic illustration of the coevolution of CRISPR repeat sequences and Cas proteins (Shah *et al.* 2010).

Recently, Makarova and colleagues (2011b) proposed the expansion of the RAMP superfamily family and its reorganization into three main groups, Cas5, Cas6 and Cas7, based on secondary and tertiary structure similarity scores and sequence conservation (HMM motifs). A single or tandem RRM motifs are found or predicted in families belonging to all three groups, but the characteristic G-rich loop is present in all families except from the Cas7 family (Makarova *et al.* 2011b)

3.1.2 The CRISPR-RAMP module (CMR)

Genes belonging to the RAMP superfamily from COGs 1769, 1336, 1367, 1604, 1337, 1332, 1567, 1583 along with genes from COGs 1353, 3337 typically form a cluster of six genes (*cmr1-6*) with conserved order, which was termed the CRISPR RAMP module (*cmr*) by Haft and colleagues (2005). This module exists in genomes in combination with other CRISPR subtypes, either physically linked or not, and is over-represented in thermophilic archaea and bacteria. The same operon was termed “polymerase-cassette module” by Makarova *et al.* in 2006, because the core COG1353 gene contains characteristic palm-domain RNA or DNA polymerase motifs (discussed later). Phylogenetic and comparative genomic analyses revealed that the *Cmr* gene cluster operates as a single genetic element and co-segregates during horizontal gene transfer events, which strongly supported the fact that these proteins would interact. Distinct hidden Markov models (HMMs) have been built for each protein family and deposited in the TIGRFAMs database (Haft *et al.* 2005). In a recent review of the various classification and nomenclature systems of the CRISPR/CAS, Makarova *et al.* (2011) proposed a new “polythetic” classification system in which the

polymerase-RAMP module proteins belong to the Type III CRISPR/Cas system. In combination with Cas1-Cas2, this system is predicted to be able to carry out both the adaptation and interference functions of the CRISPR/Cas.

The main component of the CMR module is the multidomain protein of COG1353 (Cas10). Among the identified domains is a permuted HD-superfamily hydrolase near the N-terminus, a globular uncharacterised $\alpha+\beta$ domain, a Zinc-ribbon (well-known nucleic acid interacting domain) and the core palm domain of DNA/RNA polymerases and nucleotide cyclases near the C-terminus. Sequence analysis and structure threading revealed that this core fold domain is found in reverse transcriptases, viral RNA polymerases, superfamily A & B DNA polymerases, signal-transducing adenylyl and nucleotide cyclases (termed the GGDEF domain). The conserved motifs within the extended similarity region contain several acidic residues (usually aspartates) which function as metal-coordinators in the active sites of polymerases, as well as an RRM fold motif. The co-existence of these domains and the analysis performed by Makarova *et al.* (2002) led to the suggestion that these proteins function as novel polymerases, with the N-terminal HD domains functioning either as pyrophosphatases in the context of the polymerisation reaction or as accessory nucleases.

The structure of Cmr5, the second CMR protein component not belonging to the RAMP superfamily has been solved from *Thermus thermophilus* HB8 (Tthb164, PDB: 2ZOP) (Sakamoto *et al.* 2008) and *Archaeoglobus fulgidus* DSM 4304 (Afu1862, PDB: 2OEB). The protein monomers in both cases are composed of six α -helices arranged in a barrel-like fold, with the asymmetric unit of TthCmr5 occupied by a homotrimer. The trimer surfaces present an asymmetric charge distribution, with one almost uniformly basic surface and discontinuous acidic patches on the opposite side. Within the Cmr complex, these charged surfaces could function either as protein-interaction or RNA binding platforms. Secondary and tertiary structure prediction programs such as Phyre (Kelley and Sternberg, 2009) reveal that SsoCmr5 (Sso1988) can be modeled with an estimated precision of 100% onto the structure of AfuCmr5 (Afu1862). In figure 3.3 we can observe that the whole sequence of Sso1988 can be threaded onto the AfuCmr5 structure, with the only outlier being an extended loop between residues 59-73.

Phylogenetic analysis of archaeal Cas10 homologues revealed that the type III systems (often referred to as Cmr modules in the literature) in archaea can be classified into five families A, B, C, D and E (Garrett *et al.* 2011). Subtype III-B Cmr2 family members are distributed into families B and C, subtype III-A Csm1 family representatives compose families D and E, while Csx11 family members which are part of yet unclassified type III systems form family A. In *Sulfolobus solfataricus*, we encounter two complete CMR gene clusters. The first consists of genes sso1424-

sso1432, is located between CRISPR loci B and C, and belongs to family D, indicating that this is probably a type III-A system. The second consists of genes sso1986-sso1992, is adjacent to CRISPR locus F and belongs to family B, classifying it as a type III-B system. The gene organization can be seen in figure 3.4. Two partial clusters are located between CRISPR loci C and D, consisting of genes sso1514-sso1510 and sso1725-sso1730 (without sso1728, which is a transposase), but since they lack the full set of Cmr protein components, they are considered inactive. The operons of type III systems present in *Sulfolobus solfataricus* do not contain the *cas1-cas2* gene pair. However, they most likely interact *in trans* with the *cas1* and *cas2* genes that are part of type I-A systems also present in the genome.

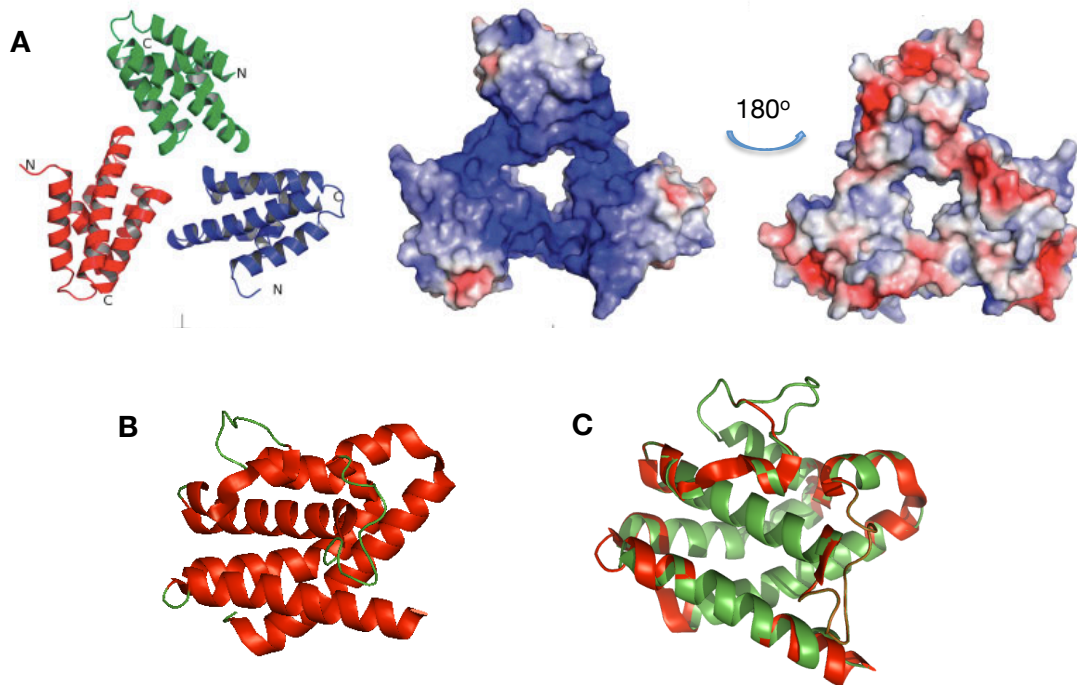


Figure 3.3: Crystal structures of Cmr5

Ribbon diagrams and electrostatic surface representations of (A) trimeric TthCmr5 (adapted from Sakamoto *et al.* 2009) and (B) monomeric AfuCmr5. (C) Superimposition of the SsoCmr5 model generated by PHYRE (green) and the AfuCmr5 (red), where we can observe the extended loop region.

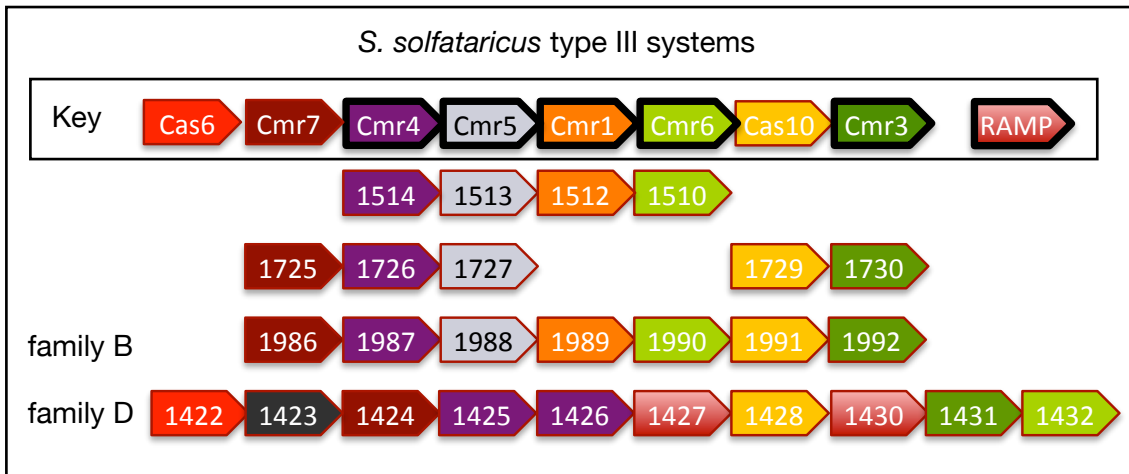


Figure 3.4: Gene organisation of the type III systems in *S. solfataricus*

Orthologous genes are colour-coded according to the Cas family key. The genes were assigned to their respective family according to their COG and their TIGR family (Makarova et al. 2011a,b). RAMP families are indicated by a black border. Genes belonging to the RAMP superfamily but not assigned to a specific family, or their classification is unresolved are coloured in light red.

Sequence analysis of the Cmr2/Cas10 protein Sso1991 (Cmr family B) reveals the conserved HD nuclease motif in the N-terminus (H15, D16), the conserved four cysteine residues (C466, C469, C514, C517) comprising the Zn-coordination domain (Zn-ribbon) and the “GGDEF” domain shared by polymerases and adenylyl cyclases (YAGGDDLL sequence, a.a.806-811) near the C-terminus. The domain organization of the protein can be seen in figure 3.5. Multiple sequence alignments illustrating the conservation of the functional motifs can be found in Appendix II. The COG1353 protein of the Cmr family D operon appears split in two in the annotated genome: Sso1428 and Sso1429, most likely due to misannotation. Nonetheless, sequence analysis reveals the presence of all three of the catalytic motifs, with the HD nuclease and Zn-ribbon motif in the N-terminus (Sso1428 sequence) and the predicted polymerase palm domain in the C-terminus (Sso1429 sequence). The third Cmr2 homologue in the genome, Sso1729, lacks the HD nuclease motif and the third cysteine which would coordinate the Zinc molecule in the Zn-ribbon motif. Although the “GGDEF” motif is intact, it would be reasonable to assume that this protein is inactive, which correlates with the degeneration of this Cmr cluster.

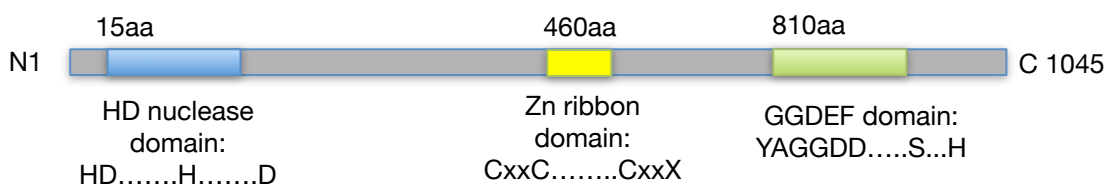


Figure 3.5: Domain organisation of SsoCmr2 (Sso1991), indicating the location of the conserved potential active sites on the sequence.

In this chapter we describe the isolation of the native Cmr family B complex from *Sulfolobus solfataricus* cell lysate, as well as our attempts to reconstruct the complex *in vitro* by recombinant expression of its individual components. Initial functional characterisation experiments were based on the bioinformatics predictions and mechanistic models and hypotheses about the CRISPR outlined in the relevant literature until the first half of 2008.

Gene annotation	CAS name	COG	TIGRfam	Protein length (aa)	MW (kD)	pI	Notes
sso1986	cmr7	_		197	22.2	7.08	
sso1987	cmr4	1336	TIGR02580	307	34.4	5.44	RAMP
sso1988	cmr5	3337	TIGR01881	156	18.2	8.66	Structure of homologue available
sso1989	cmr1	1367	TIGR01894	467	54.9	8.32	RAMP
sso1990	cmr6	1604	TIGR01898	284	32.5	8.8	RAMP
sso1991	cmr2/cas10	1353	TIGR02577	1045	121.69	7.95	HD domain, Zn-ribbon, palm domain
sso1992	cmr3	1769	TIGR01888	314	36.19	5.68	RAMP

Table 3.1: The Cmr family B cluster in *S. solfataricus*.

3.2 Expression and purification of recombinant Cmr7

The gene encoding Cmr7 (sso1986) was amplified by PCR from *S. solfataricus* genomic DNA and successfully cloned into the pEHISTEV vector to enable expression of the recombinant protein with an N-terminal polyhistidine tag. The construct was sequenced to verify its integrity and lack of mutations and transformed into the *E. coli* expression host strain BL21 (DE3). Protein was expressed as described in chapter 2 and purified to homogeneity with a two-step purification scheme which included affinity (nickel-chelating) and size-exclusion chromatography, from which it eluted as a monomer. The N-terminal polyhistidine tag was cleaved by incubation with the TEV protease overnight at room temperature, as described in Materials and Methods. Protein identity was confirmed by mass spectrometry.

The expression levels of recombinant Cmr7 were very high and the protein was highly soluble with only a minimum amount of the total expressed protein lost in the insoluble cell fraction, as illustrated in figure 3.6. This can be attributed to aggregation due to the high amounts of the expressed protein in the cell. The protein was purified to homogeneity with a final recovery of 14 mg protein per litre of culture, 1 mg of which was used for the production of polyclonal sheep antibodies. The outcome of each purification step can be seen in figure 3.6. The proteins' estimated molecular

weight from SDS-PAGE analysis is in agreement with the calculated molecular weight of 22,151 Da. The protein eluted as a single monodispersed peak from gel filtration chromatography, which is a prerequisite to continue with crystallographic studies of the protein.

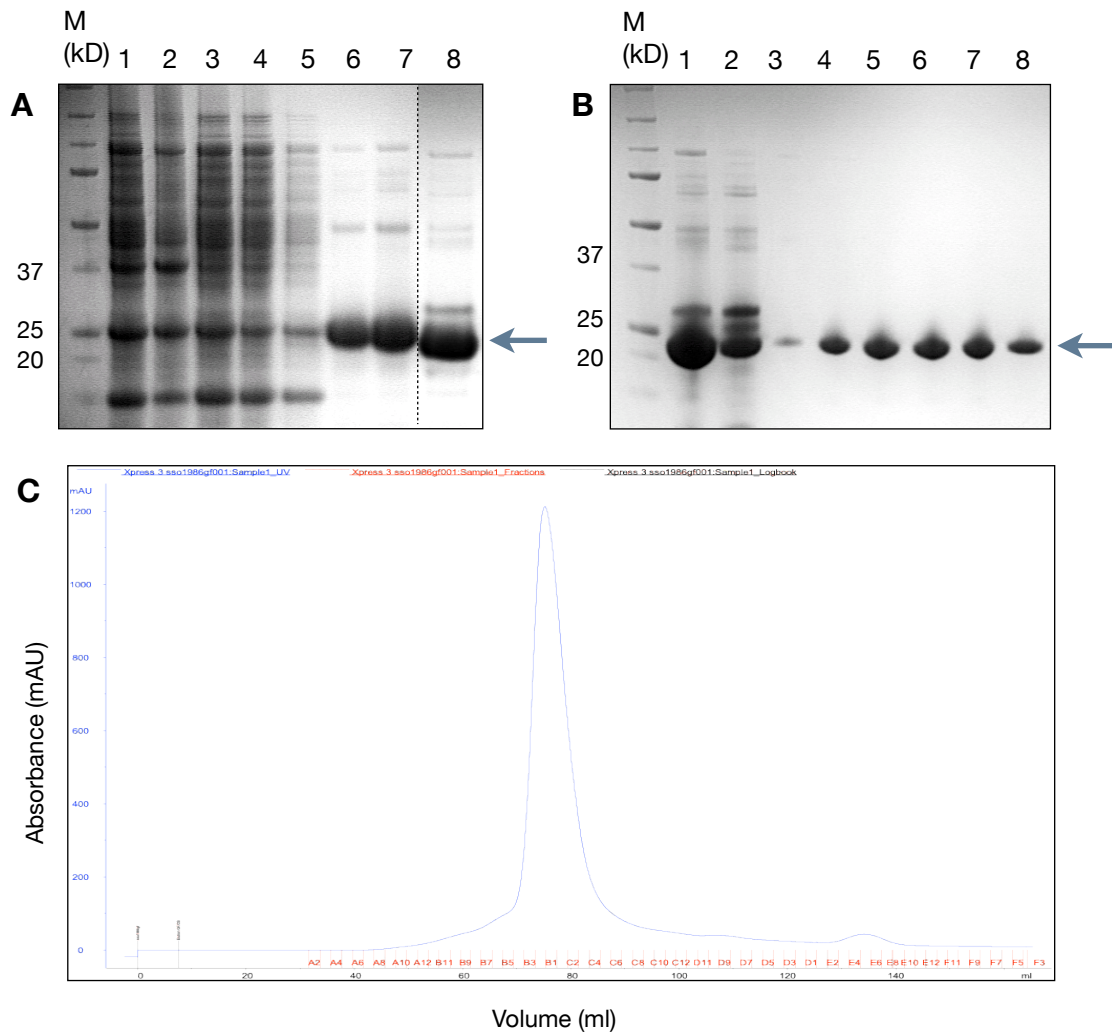


Figure 3.6: Purification of Cmr7

Purification stages monitored by SDS-PAGE. Lanes (A): M, protein size marker; 1, whole cell extract; 2, insoluble fraction; 3, soluble fraction; 4, flow-through; 5, wash; 6-7, Cmr7 elution from first Ni-affinity column; 8, protein sample after TEV cleavage; Dashed line indicates non-contiguous lanes in gel (B): M, protein size marker; 1-2, flow-through from second Ni-affinity column containing cleaved Cmr7; 3-8 peak fractions from gel filtration column containing pure Cmr7. (C): Chromatogram of Cmr7 purification by gel filtration chromatography.

3.3 Crystallographic study of Cmr7

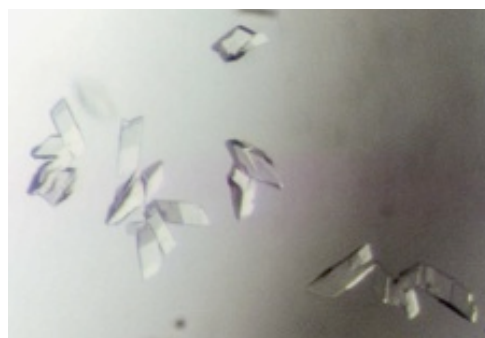
3.3.1 Crystallisation and structure solution of Cmr7

The crystallisation and structure solution of Cmr7 was carried out in collaboration with the SSPF (Oke *et al.* 2011). The Pre-Crystallisation Test (Hampton Research) for Cmr7 indicated that the optimal concentration for crystallisation screens was 10 mg/ml. The Classics and JCSG crystal screens were set up manually as described in chapter 2 and after 3 days microcrystals were observed in 1 M lithium

chloride, 0.1 M trisodium citrate pH 4 and 20 % PEG 6000 (JCSG). Stochastic optimisation screens were designed around this condition with the assistance of Dr Kenneth Johnson, varying the following parameters: the precipitant type and concentration (PEG 4000, 6000, 8000), the salt (LiCl, LiS) and its concentration, the buffer pH (pH 4, 4.5, 5), the concentration of additives such as glycerol and ethylene glycol and the protein concentration (10 mg/ml, 5 mg/ml) in order to enable the formation of single, well-ordered crystals by delaying nucleation.

The growth of small, bipyramidal crystals was observed in 10mg/ml protein, 19.2% PEG 6000, 0.1 M sodium citrate pH 4, 0.93 M LiCl and 9.5 % ethylene glycol (figure 3.7). Crystals were cryo-protected in 21.0 % PEG 6000, 0.1 M Na-Citrate pH 4.0, 0.5 M LiCl, 25 % ethylene glycol, frozen in liquid nitrogen and sent to the European Synchrotron Radiation Facility (ESRF, Grenoble) for diffraction data collection. Crystals for SAD phasing were soaked in 200 mM Me_3PbCl , 500 mM K_2OsO_4 and 200 mM K_2OsCl_6 . The collected dataset had a resolution of 2.05 Å and the structure was solved and refined by the SSPF as described in Oke *et al.* (2010). The final structure coordinates and refinement statistics have been deposited in the protein data bank with accession code 2X5Q.

Figure 3.7 : Crystals of Cmr7, grown under the conditions mentioned in the text



3.3.2 Structure of Cmr7

Two monomers of Cmr7 (Sso1986) were found in the asymmetric unit of the crystal, suggesting that the protein is a dimer. The final model consists of 184 amino acid residues, arranged in 4 α -helices and 15 β -strands. Residues 1-7 and 125-131 are missing from the final model as the electron density was not defined. The structure of the Cmr7 protomer reveals a two-domain $\alpha+\beta$ architecture, with a small helical domain connected to a larger all- β domain. The latter is comprised by three twisted antiparallel β -sheets, formed by strands $\beta 2-\beta 3-\beta 4-\beta 10$ (sheet 1), $\beta 8-\beta 9-\beta 16-\beta 15$ (sheet 2) and $\beta 1-\beta 10-\beta 7$ (sheet 3), while the $\beta 5-\beta 6$ hairpin interacts with the helical bundle formed by $\alpha 2-\alpha 4$ (figure 3.8 A, C). Two monomers interact via the β -sheet domains to form the dimer (figure 3.9 A). The electrostatic surface calculations by the APBS software package (Baker *et al.* 2001) reveal a positive surface wrapping around the protein, and a smaller region of negative charge concentrated on one side, in what

appears to be a small “tunnel” (figure 3.8 B). The two monomers exhibit slight structural variation, indicating a certain degree of conformational flexibility.

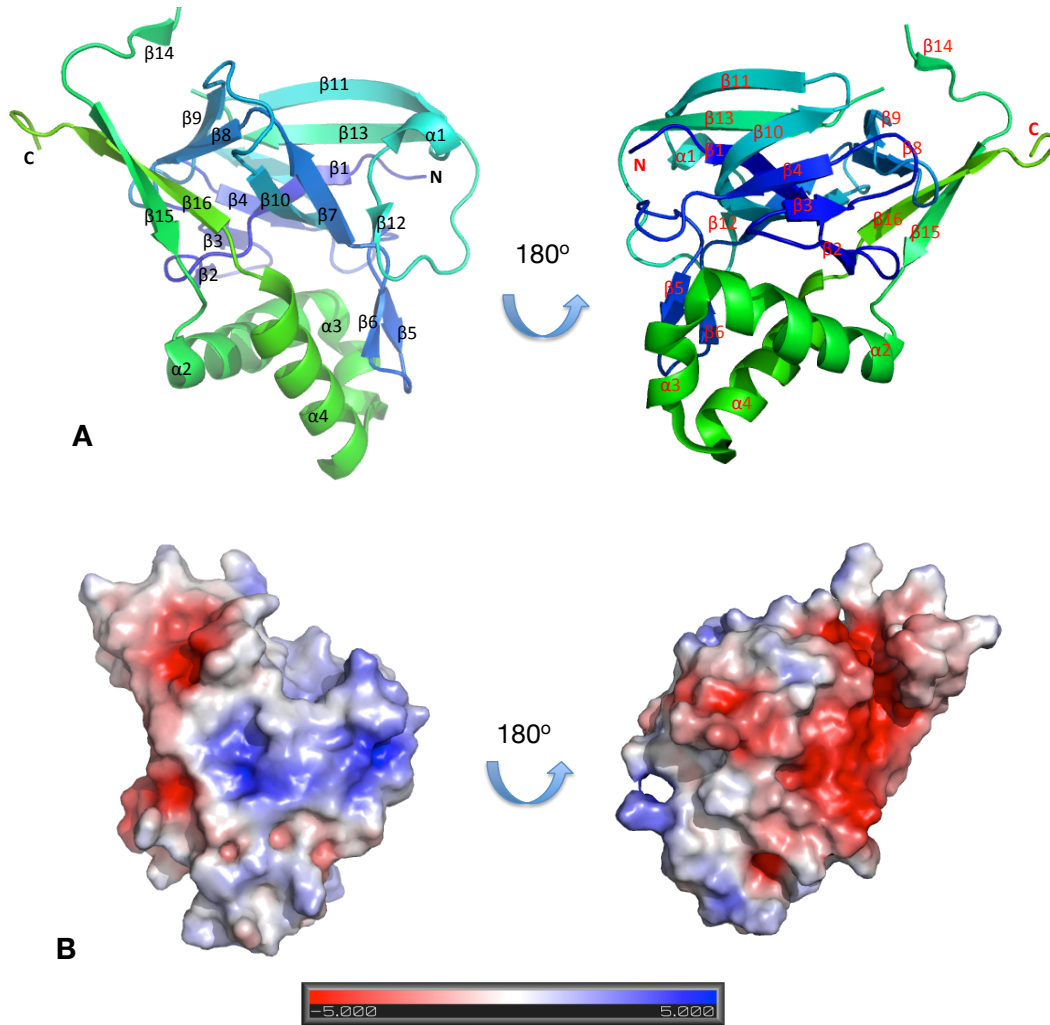
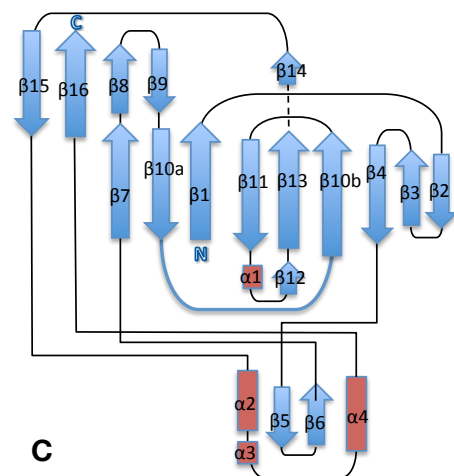


Figure 3.8: Structure of the Cmr7 monomer
 (A) Cartoon diagram of the Cmr7 monomer. The secondary structure elements are numbered as in (C) Figures generated with MacPymol. (B) Electrostatic surface of the Cmr7 monomer generated with APBS tools, oriented as in (A). (C) Topology diagram of Cmr7, illustrating the connectivity of the various secondary structure elements.



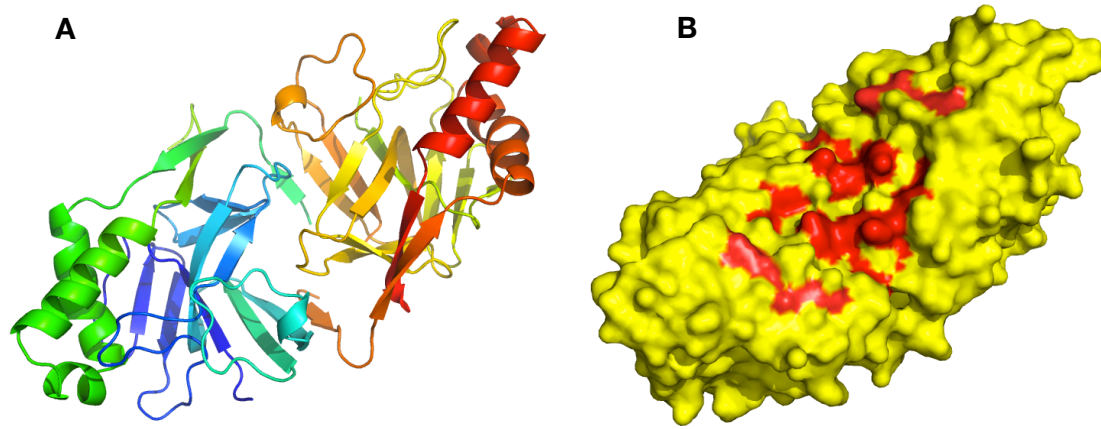


Figure 3.9: Cmr7 dimer and conserved surfaces

(A) Cartoon diagram of the Cmr7 dimer. The two chains are colored in orange-red (chain A) and blue-green (chain B). (B) Surface representation of the Cmr7 dimer illustrating the location of conserved residues (red). Orientation as in (A).

Structural similarity searches using DALI (Holm and Rosenstrom, 2010) did not reveal any significant structural neighbours, indicating that it is a novel fold. This was expected since previous sequence analysis failed to detect any conserved motifs or classify this sequence to any known protein family, and located homologues of Cmr7 only in Sulfolobales. A series of residues forming the first half of strand β 10 seem to be strictly conserved among all Cmr7 homologues, which along with a conserved Arg-Lys pair seem to cluster on the positive surface side of the molecule (figure 3.9 B). No functional predictions can be made for this Cmr subunit at this point.

3.4 Expression and purification of recombinant Cmr4

The gene encoding Cmr4 (*sso1987*) was amplified by PCR and cloned into the pDEST14 expression vector as described in Materials and Methods in order to enable N-histidine-tagged protein expression. The construct was fully sequenced to verify the absence of mutations and protein was expressed in *E. coli* C43 (DE3) as described in Materials and Methods. Purification involved affinity (nickel-chelating) chromatography, TEV cleavage of the polyhistidine tag and a final polishing step with size exclusion chromatography (figure 3.10). The protein identity was confirmed by mass spectrometry. All steps were carried out on ice due to protein precipitation at elevated temperatures.

Protein expression levels were moderate and a significant portion of the expressed protein was found in the insoluble fraction. One possible explanation is inadequate cell lysis, but a more probable one is that the protein would require its physiological partners to enhance its stability and solubility in solution. Typical yields ranged from 0.19 - 0.26 mg/L of culture, with predictable losses during concentration

and chromatography steps. The calculated molecular weight of the protein as confirmed on SDS-PAGE is 34.7 kDa. Analytical gel filtration performed on a Superdex 200 10-300 column (GE Healthcare) by the SSPF indicated that the Cmr4 migrated as a trimer in 150 mM NaCl, although this is not necessarily the protein's oligomeric state within the Cmr complex.

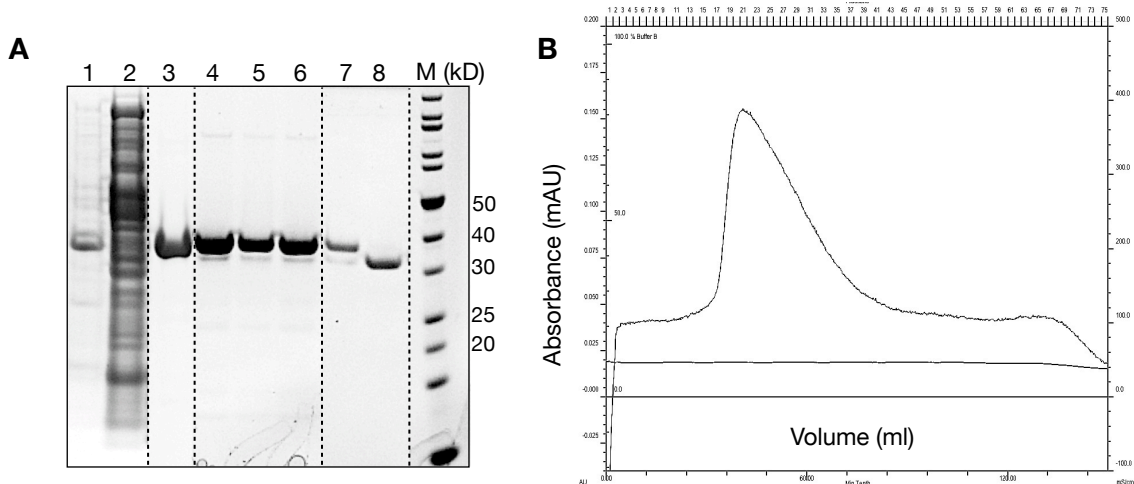


Figure 3.10: Purification of Cmr4

(A): Stages of purification on SDS-PAGE. 1, insoluble cell fraction; 2, soluble cell fraction; 3, Cmr4 eluting from first nickel-chelating affinity chromatography; 4-6, peak fractions from gel filtration chromatography; 7, sample before TEV cleavage of six-histidine tag; 8, pure Cmr4 after TEV cleavage; M, protein size marker. Dashed lines indicate non-contiguous lanes in different gels. (B) Chromatogram of Cmr4 purification by gel filtration chromatography.

3.5 Expression and purification of recombinant Cmr1

The *S. solfataricus cmr1* gene (sso1989) was PCR amplified and cloned into the pDEST14 expression vector as described in Materials and Methods. The construct was sequenced to confirm it was free of base pair mutations and protein was expressed in host *E. coli* strain C43 (DE3). The protein was purified by a two-step purification scheme, consisting of affinity (nickel-chelating) and size exclusion chromatography. After the first affinity column, the protein-containing sample was subjected to a 2 hour incubation with TEV protease at room temperature and passed through a second affinity column in order to separate tagged from untagged protein. Protein identity was verified by mass spectrometry. The protein-containing samples were kept on ice due to protein instability at higher temperatures. An overview of the purification can be seen in figure 3.11.

Expression levels were moderate with a final yield of 0.4 -1 mg/L culture. Only a fraction of the total expressed protein was soluble, a situation to be expected as the Cmr proteins in vivo form part of a larger complex and would not be expressed on their own. The observed molecular weight of Cmr1 on SDS-PAGE analysis confirmed the calculated molecular weight of 54.8 kDa. The protein migrated as a monomer

during analytical gel filtration performed by the SSPF on a Superdex 200 10-300 column, but as explained for Cmr4 this does not necessarily reflect the protein's state within the Cmr complex.

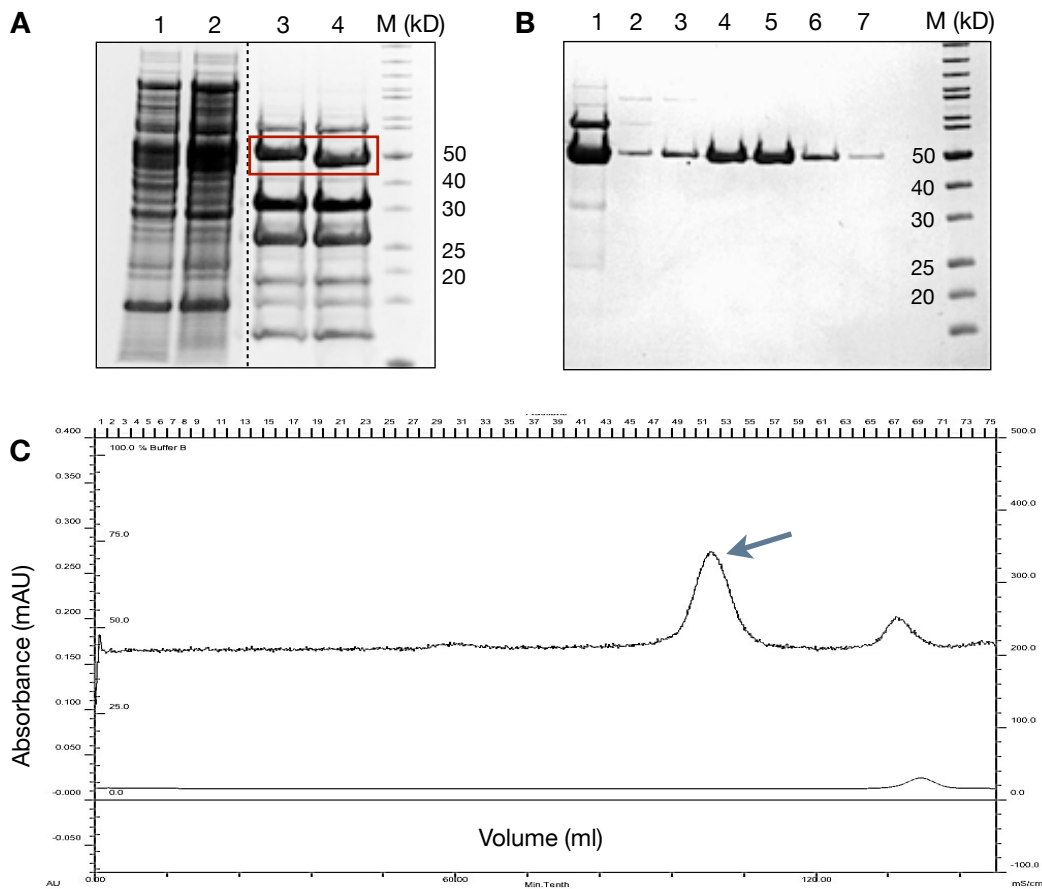


Figure 3.11: Purification of Cmr1

Purification stages monitored by SDS-PAGE. Lanes (A): 1, whole cell extract; 2, soluble fraction; 3, Cmr1 elution from first Ni-affinity column; 4, protein sample after TEV cleavage; M, protein size marker; (B): 1, flow-through from second Ni-affinity column containing cleaved Cmr1; 2-7 peak fractions from gel filtration column containing pure Cmr1; M, protein size marker. Dashed lines indicate non-contiguous lanes in different gels. (C): Chromatogram of Cmr1 purification by gel filtration chromatography.

3.6 Protein interactions between recombinant CMR components

With the addition of recombinant Cmr3 (Sso1992), expressed and purified by the SSPF, all the aforementioned recombinant Cmr proteins were obtained in both tagged and untagged form. This enabled them to be assayed in pairs for *in vitro* interaction using Ni-NTA magnetic agarose beads as described in Materials and Methods (figure 3.12). This approach would enable us to study the topography of the native CMR protein complex by reconstituting the pairwise interactions involved.

The following combinations were tested:

- His-tagged Cmr3 with Cmr7 (figure 3.13 B)
- His-tagged Cmr3 with Cmr4 (figure 3.13 B)
- His-tagged Cmr3 with Cmr1
- His-tagged Cmr1 with Cmr7 (figure 3.13 A)
- His-tagged Cmr1 with Cmr4 (figure 3.13 A)
- His-tagged Cmr3 with Cmr1, Cmr4 and Cmr7 combined (figure 3.13 C)

Tagged proteins were bound to the nickel-loaded beads and then incubated for 1 hour at room temperature with the appropriate partner. After thorough washing at low salt (100 mM NaCl) and imidazole (10 mM) concentrations the bound proteins were eluted from the beads with 500 mM NaCl, 500 mM imidazole and analysed on SDS-PAGE. Appropriate controls were carried out to test for binding of the untagged proteins to the magnetic beads. Of all pairs tested, only Cmr1 and Cmr3 were found in the eluate fraction, indicating a direct physical interaction, although the stoichiometry of the interaction cannot be determined from such a qualitative experiment. Whether the formation of this complex would provide an initial scaffold for the binding of the Cmr4 and Cmr7 was tested in the last experiment (figure 3.13 C), where all the available recombinant proteins were pre-incubated for 1hr at room temperature prior to binding to the nickel-coated beads. In all experiments, the amount of each protein was kept at 7.5 μ g, which was the maximum bead capacity.

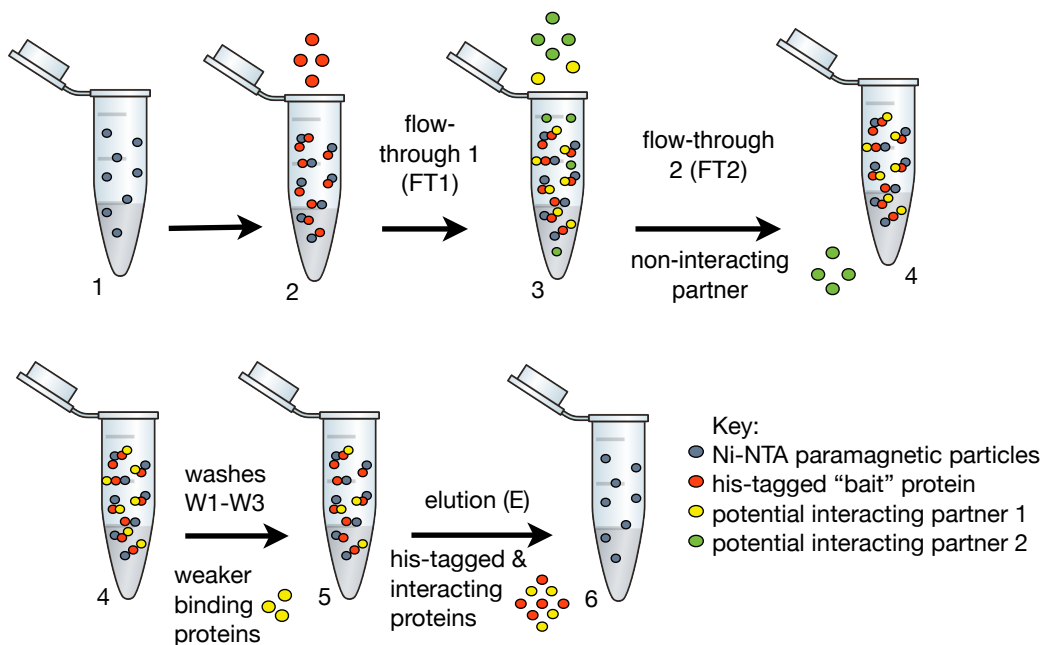


Figure 3.12: Experimental setup for detecting protein-protein interactions using Ni-loaded magnetic beads

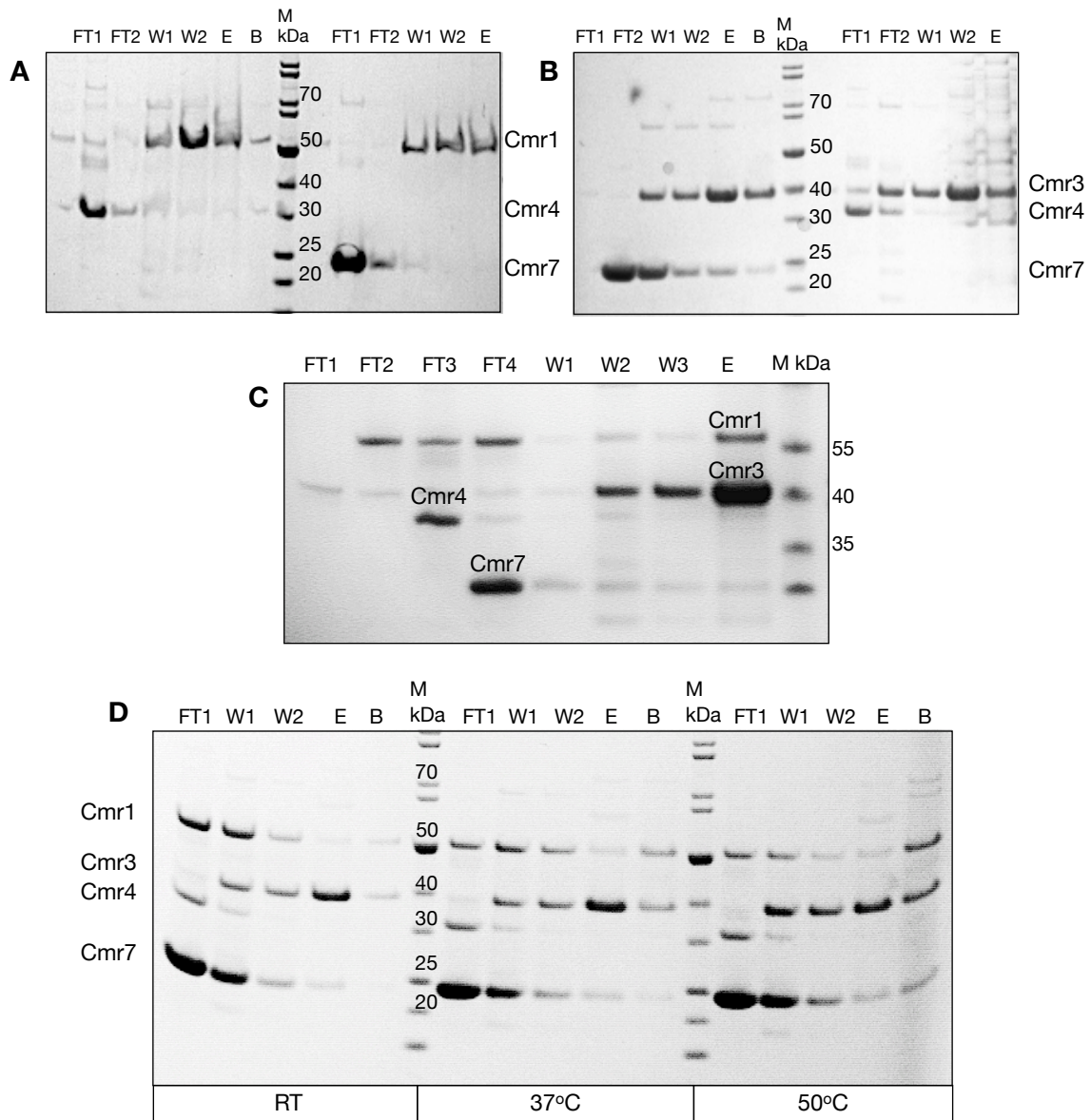


Figure 3.13: Interactions between recombinant Cmr subunits

(A) Pairwise interactions between his-tagged Cmr1/native Cmr4 (first 6 lanes) and his-tagged Cmr1/native Cmr7 (last 6 lanes). The collected fractions are labelled as (FT1), flow-through after the his-tagged protein was added to the beads; (FT2) flow-through after the interacting partner was added; (W1-2) washes with increasing imidazole concentration; (E) elution with 500mM imidazole; (M) protein size marker.

(B) Pairwise interactions between his-tagged Cmr3/native Cmr7 (first 6 lanes) and his-tagged Cmr3/native Cmr4 (last 6 lanes). Fractions labelled as for (A). No interactions were observed in either combination.

(C) Interaction between his-tagged Cmr3 and native Cmr1, Cmr4 and Cmr7, combined. Proteins were added sequentially and the fractions are labelled as (FT1), flow-through after the his-tagged protein was added to the beads; (FT2) flow-through after addition of Cmr1; (FT3) flow-through after addition of Cmr4; (FT4) flow-through after addition of Cmr7; (W1-3) washes of increasing NaCl and imidazole concentration; (E) elution with 500 mM NaCl and imidazole; (M) protein size marker. Cmr1 and Cmr3 are clearly shown to interact.

(D) Repetition of the experiment depicted in (C) investigating the effect of temperature and CRISPR DNA on the interaction. An identical image was obtained with the same experiments in the absence of CRISPR DNA. Fractions labelled as before with the addition of (B), sample incubated at 90°C in SDS-PAGE loading buffer. Interaction between Cmr1 and Cmr3 is clearly observed in lane (B) at 37°C and 50°C.

No additional interactions were observed under the conditions tested, suggesting that other subunits and possibly appropriate CRISPR-related nucleic acid fragments play a structural role. Moreover, temperature-related conformational changes could facilitate interactions, considering that the physiological temperature range for *S. solfataricus* is 70°C - 80°C. In order to investigate these possibilities, the experiment was repeated in the presence of CRISPR DNA sequences (CRISPR locus construct I, see Appendix I) and in three different temperatures (room temperature ~21°C, 37°C and 50°C). In the assay depicted in figure 3.13 D, 50 ng of CRISPR DNA were added to a 25 µl protein mixture of his-tagged Cmr3 and non-tagged Cmr1, Cmr4, and Cmr7 (7.5 µg of each) and incubated for 30 min at different temperatures. The protein mixture was then added to the nickel-coated beads and processed as described in Materials and Methods. The presence of CRISPR DNA did not seem to mediate additional interactions between the Cmr subunits, although the already observed interaction between Cmr1 and Cmr3 became more pronounced at elevated temperatures compared to room temperature. A small amount of Cmr7 seems to be carried over the elution fractions at 50°C but it was not considered significant as the majority of the protein was found in the flow-through.

3.7 Nucleic acid binding by recombinant CMR proteins

To assess the nucleic acid binding capabilities of the recombinant CMR proteins we performed gel mobility shift assays with various DNA or RNA substrates. In order to determine whether the protein exhibits sequence specificity, RNA oligonucleotides corresponding to the CRISPR repeat sequences of the CRISPR loci of *S. solfataricus* were incubated with serial dilutions of the proteins at room temperature or 50°C prior to separation by native polyacrylamide gel electrophoresis. The substrates also included unspecific (20-mer and 45-mer ss/ds poly-T/AT DNA, 20-mer poly-U RNA) sequences to investigate unspecific nucleic acid binding capacity. All the substrate sequences can be found in chapter 2, paragraph 2.6.1. Representative gels from multiple experiments are shown in figure 3.14.

Only Cmr1 was clearly shown to bind single strand unspecific and CRISPR repeat RNA when left to interact for 20 min at 37°C, as illustrated in figure 3.14, third column, rows C & D. A preference for CRISPR-related RNA is observed which could be attributed to a theoretical secondary structure adopted by the RNA fragment, although an apparent dissociation constant (K_D) could not be estimated as the protein could not be purified at a high enough concentration to bind all the available substrate and cause a complete shift on the gel. It was also investigated whether the interaction between Cmr1 and Cmr3 could alter the RNA binding ability of Cmr1 or whether the presence of the other recombinant CMR proteins (Cmr4 and Cmr7) also had an effect.

None of these approaches resulted in an observable change in the binding ability of Cmr1 (figure 3.15).

None of the other recombinant Cmr proteins exhibited nucleic acid binding capacities with the substrates and under the conditions tested here. Results from this EMSA studies should not be considered definitive, as there are many limitations. They fail to detect transient and dynamic protein interactions, as non-stable protein-DNA complexes dissociate while migrating through the gel matrix. This could lead to underestimation of the apparent K_d or non-detection of the interaction.

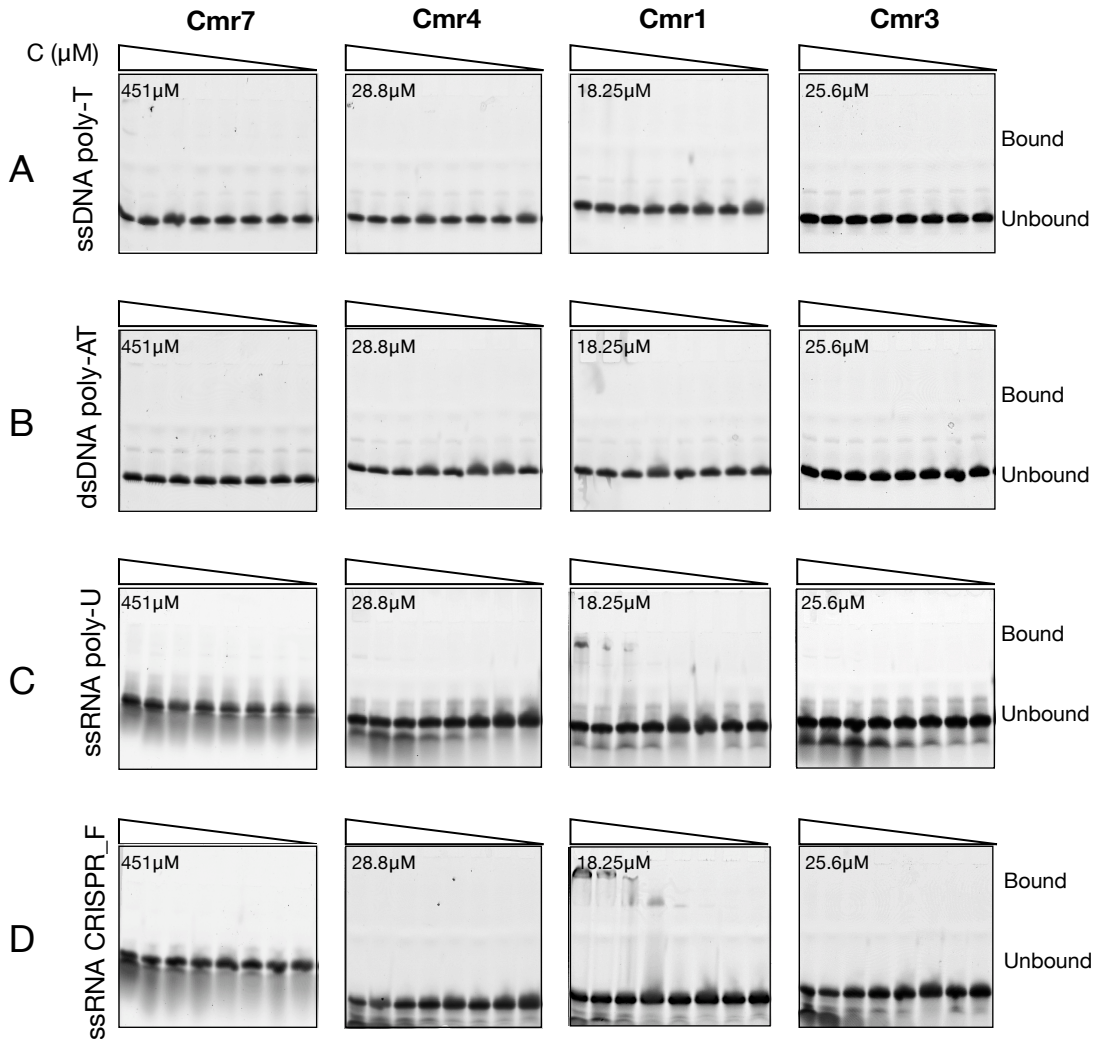


Figure 3.14: Electrophoretic Mobility Shift Assays investigating the binding affinity of recombinant Cmr proteins to various nucleic acid substrates, carried out in collaboration with Paul Talbot.

Reactions are organised in rows according to substrate: (A) 20-mer poly-T ssDNA; (B) 45-mer polyAT dsDNA; (C) 20mer-poly-U ssRNA; (D) 25mer CRISPR_F ssRNA; and in columns according to the protein tested. In each assay two-fold serial dilutions of protein were incubated with $1 \mu\text{M}$ of fluorescein-labelled substrate for 20 min at 37°C . Initial concentrations are indicated in the top left corner of each panel. Concentrations for each protein are from highest to lowest, in μM : Cmr7: 451, 225.5, 112.75, 56.37, 28, 14, 7, 3.5; Cmr4: 28.8, 14.4, 7.2, 3.6, 1.8, 0.9, 0.45, 0.225; Cmr1: 18.25, 9.125, 4.56, 2.28, 1.14, 0.57, 0.285, 0.142; Cmr3: 25.6, 12.8, 6.4, 3.2, 1.6, 0.8, 0.4, 0.2. In the third gel of rows C and D we can observe a shift in the highest concentrations as a result of binding of Cmr1 to ssRNA.

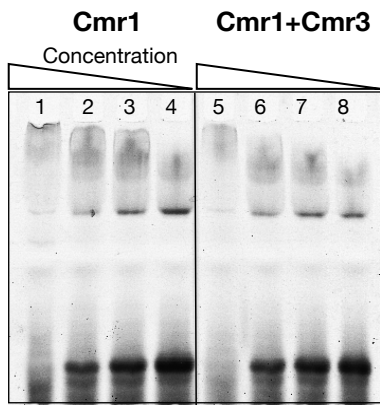


Figure 3.15: Effect of Cmr3 on Cmr1 RNA binding.

In lanes 1-4, serial dilutions of Cmr1 (in μM : 9.125, 4.5, 2.3, 1.14) were incubated with $1\mu\text{M}$ of CRISPR_F ssRNA for 20 min at 37°C . In lanes 5-6, equimolar amounts of Cmr3 were added to each reaction. No change in the binding affinity was observed.

3.8 Isolation of the native CMR complex from *Sulfolobus solfataricus*

3.8.1 Antibody assisted purification of the SsoCMR complex

A purification scheme based on the structural stability, size and physicochemical properties of the complex was designed, including four successive chromatographic steps and antibody-assisted identification of the complex after each step. The experimental procedure described here was further optimized in the lab by Paul Talbot to produce a higher yield and a high purity sample of the SsoCMR complex (figure 3.20). The high expression levels, solubility and stability of recombinant Cmr7 made it an obvious target to raise antibodies against, which could be then used to track the complex along the purification pipeline.

Step 1: Affinity chromatography

S. solfataricus P2 cell lysate was prepared from 10-50 gr of cell pellet by standard procedure described in Materials and Methods. The soluble fraction was passed through a syringe-driven sterile $0.45\ \mu\text{m}$ filter and subjected first to affinity chromatography on a 5 ml HiTrap Heparin HP column (GE Healthcare), equilibrated with 20 mM MES pH 6, 1 mM EDTA, 1 mM DTT (figure 3.16 A). In this context, this type of affinity chromatography makes use of the fact that heparin (being a sulphated glycosaminoglycan) mimics the polyanionic structure of nucleic acids, enabling the separation of nucleic acid - interacting molecules from the rest. Bound proteins were eluted over a 0 - 1 M NaCl gradient in 4 ml fractions and stored on ice until needed. Collected fractions were blotted on nitrocellulose membrane as described in Materials and Methods and anti-Cmr7 antibodies were used to detect fractions containing the CMR complex (figure 3.16 B). The respective fractions were pooled together and applied to the next column. As can be seen in figure 3.16 A, the CMR complex eluted at approximately 750 mM NaCl, indicating a tight interaction with the column and therefore a strong nucleic acid binding capability.

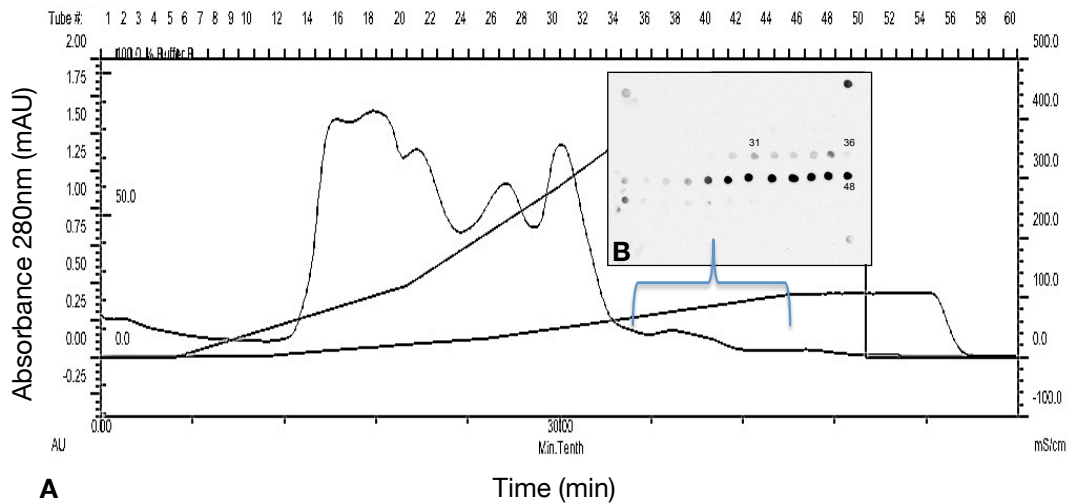


Figure 3.16: First step of SsoCMR purification by affinity chromatography.

(A) Chromatogram of the elution profile of HiTrap Heparin HP. The dot blot of the elution fractions can be seen figure (B), where the fractions containing Cmr7 have interacted with the anti-Cmr7 antibody and given a sharp signal. The positive fractions correspond to the indicated area in the chromatogram, where a late elution peak is observed. Fractions 37-49 were combined and subjected to the next purification step.

Step 2: Size exclusion chromatography

The CMR-containing fractions pooled from the heparin column were concentrated to an appropriate volume in a Vivaspin concentrator with 10 kDa cutoff limit, and loaded on a HiLoad 26/60 Superdex 200 gel filtration column (GE Healthcare), equilibrated with 20 mM MES pH 6, 200 mM NaCl, 1 mM EDTA, 1 mM DTT. Fractions of 4 ml were collected and dot blotted with anti-Cmr7 antibodies as before. As observed in figure 3.17, the complex eluted in a broad peak containing other high molecular weight molecules, in agreement with the size and stability of the structure. A more accurate size for the complex could not be estimated at this step, since the sample contains still many cell contaminants.

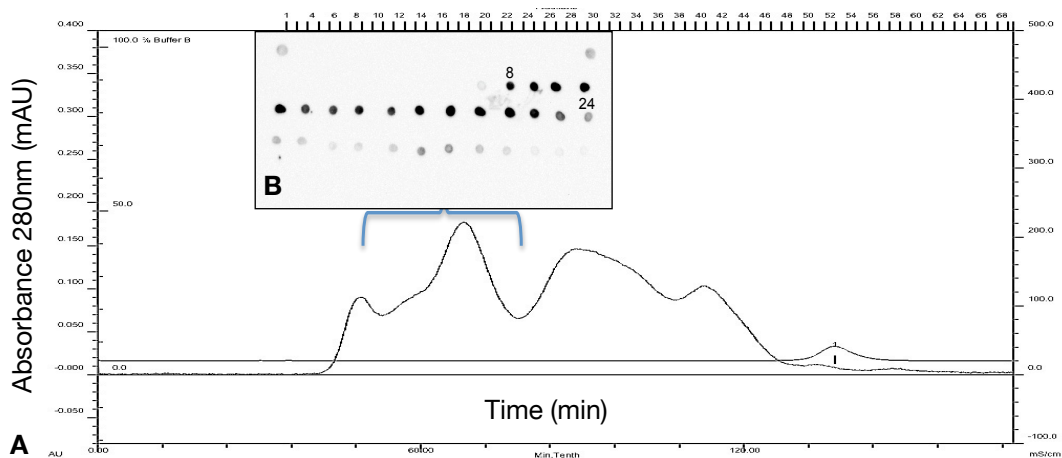


Figure 3.17: Second step of SsoCMR purification by size exclusion chromatography.

(A) Gel filtration elution profile. (B) Dot blot of the collected fractions where we can observe that the positive fractions correspond to an early eluting peak, presumably due to the large size of the complex.

Step 3: Cation exchange chromatography

The pooled fractions from gel-filtration chromatography were buffer exchanged to low salt buffer and subsequently applied to a 1 ml MonoS 5/50 GL (GE Healthcare) cation exchange column, equilibrated with 20 mM MES pH 6, 1 mM EDTA, 1 mM DTT. Bound proteins were eluted over a 0 - 1 M NaCl gradient in 1 ml fractions, and dot-blotted as described. The complex eluted approximately at 250 mM NaCl (figure 3.18), indicating a rather weak or unevenly distributed positive surface charge.

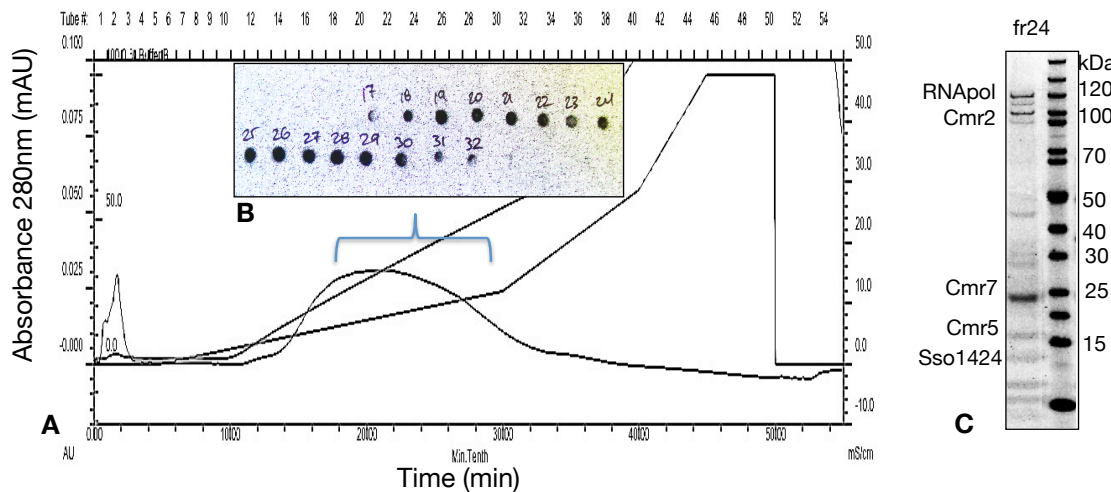


Figure 3.18: Third step of SsoCMR purification by cation exchange chromatography.

(A) Elution profile of 1 ml MonoS 5/50 GL with a 0 - 1 M NaCl gradient. (B) Blotted fractions giving a positive signal when incubated with anti-Cmr7 antibody. Fragments of this homogeneous peak were analysed by SDS-PAGE and Ruby Sypro staining, and protein identification was carried out by mass spectrometry. A representative fraction of the peak can be seen in panel (C), where we observe a complex mixture of bands, with both the CMR complexes and RNA polymerase subunits present, as detected by mass spectrometry.

Step 4: Anion exchange chromatography

As a final purification step, a 1 ml MonoQ 5/50 GL (GE Healthcare) anion exchange column was employed. The pooled sample from the previous step was concentrated to an appropriate volume and buffer exchanged to 20 mM Tris-HCl pH 8, 1 mM EDTA, 1 mM DTT prior to loading on the MonoQ column, equilibrated in the same buffer. Bound proteins were eluted over a 0 - 1 M NaCl gradient in 1 ml fractions, in which the presence of the CMR complex was detected by dot blot with anti-Cmr7 antibodies as before (figure 3.19). The complex eluted at approximately 350mM NaCl, reflecting a weak or unevenly distributed negative surface charge. SDS-PAGE analysis revealed that the two distinct peaks observed in the chromatogram correspond to the separation of the 12-subunit SsoRNA polymerase complex (sharp peak) from the SsoCMR complex (broader peak), which seems to contain only a few remaining contaminants (figure 3.19, D). The stoichiometry of the complex could not be determined at this stage, but it could be roughly estimated that there are equimolar amounts of Cmr1-6 and an excess of Cmr7. The complex-containing fractions were pooled together, concentrated at approximately 100 μ g/ml

and stored at -80°C until required. Mass spectrometry analysis of the protein content of the samples will be discussed in the following paragraph.

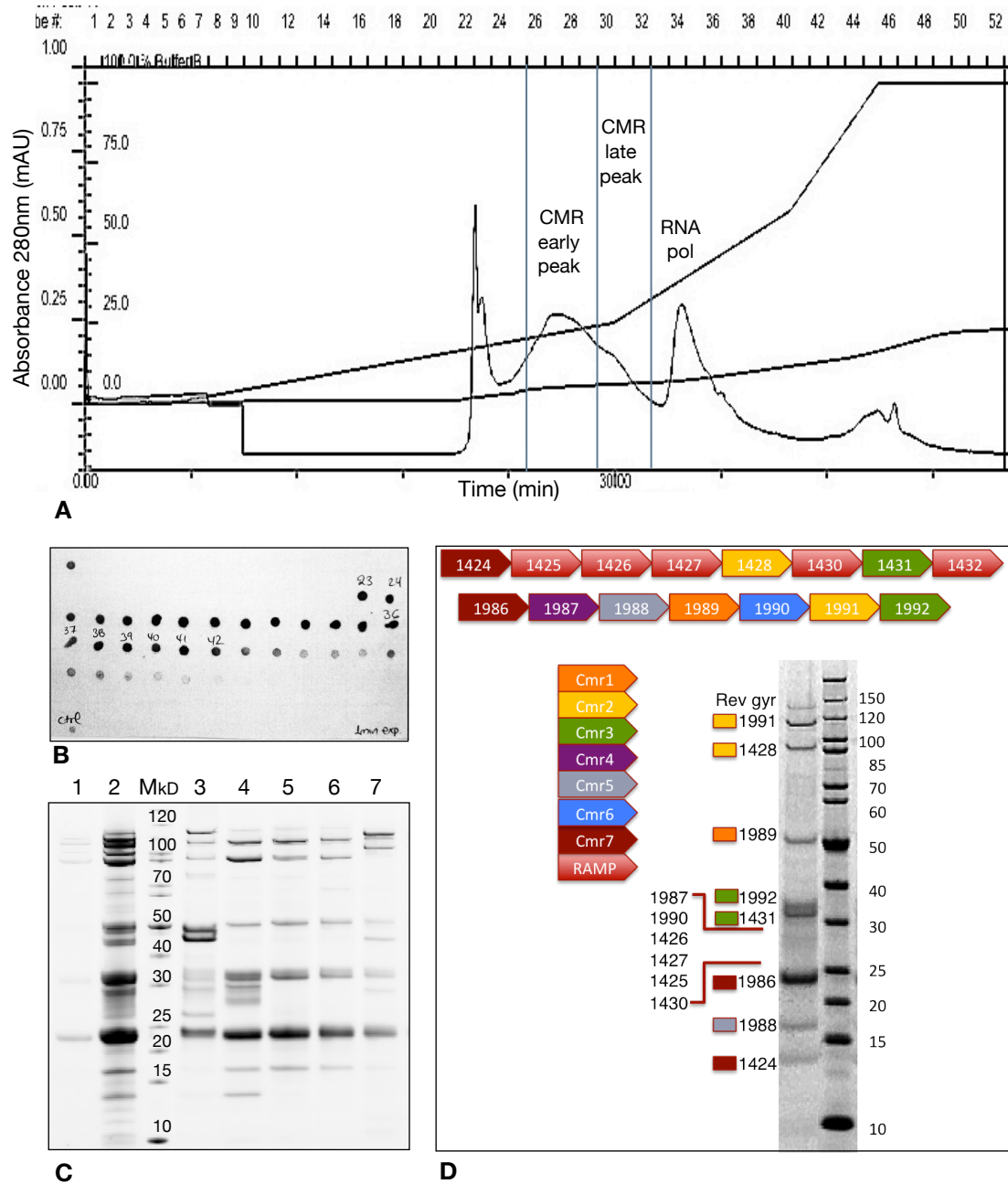


Figure 3.19: Fourth step of SsoCMR purification by anion exchange chromatography.

(A) Elution profile of the 1ml MonoQ 5/50 GL with a 0 - 1M NaCl gradient. The broad CMR elution peak is divided into two sections indicated on the chromatogram and analysed by mass spectrometry and SDS-PAGE (C). In this stage the CMR complex separates from the RNA polymerase complex which elutes later from the column in a sharp peak. (B) Dot blot of the MonoQ elution indicating the Cmr7-containing fractions, which initially seem to cover a broad area between fr. 23-42. (C) When analysed by SDS-PAGE and Ruby Sypro staining, the peak fractions (lanes 3-6) reveal a heterogeneous protein composition and were subsequently analysed by mass spec (see paragraph 3.8.2). Lane description: 1, Gel Filtration CMR pool; 2, MonoS CMR pool; M, protein size marker (kD); 3, MonoQ fraction 23; 4, MonoQ fr27; 5, MonoQ fr30; 6, MonoQ fr32; 7, MonoQ fr34 RNA polymerase complex. (D) Schematic description of the protein composition of the final CMR complex pool enabled by mass spectrometry analysis. Component proteins are colour coded and defined by gene name and module type. Both SsoCMR complexes are present.

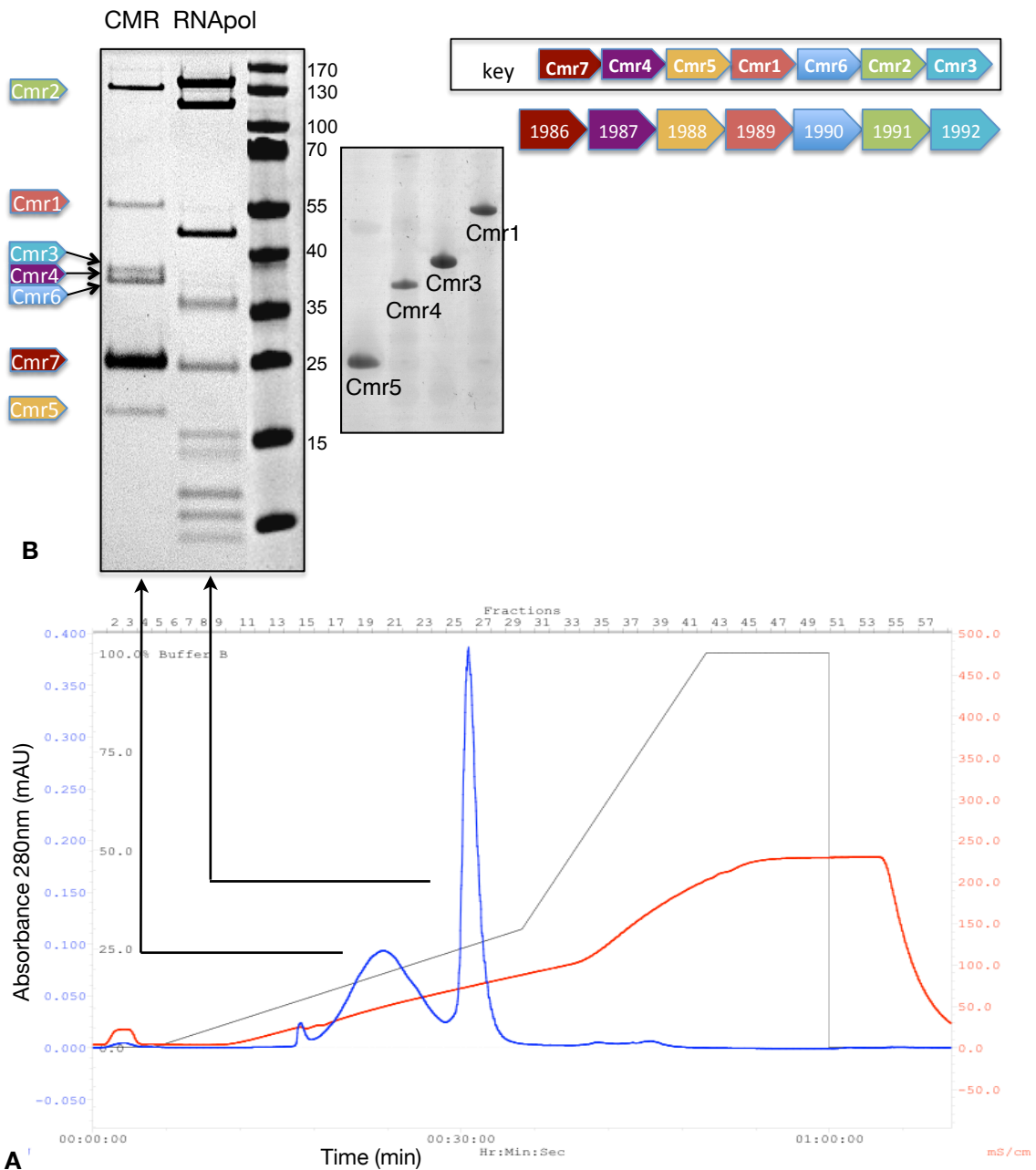


Figure 3.20: Optimisation of the SsoCMR purification procedure by Paul Talbot with a larger starting culture.

(A) Elution profile of the 1 ml MonoQ 5/50 GL with a 0 - 1 M NaCl gradient. (B) SDS-PAGE analysis of the peak fractions. The SsoCmr family B complex elutes as a single monodispersed peak and is isolated from the SsoCmr family D cluster, as confirmed by mass spectrometry (data not shown). The second sharp peak contains in an almost pure form the SsoRNA polymerase cluster. The adjacent gel image shows the purified recombinant CMR proteins for comparison.

3.8.2 Identification of the CMR complex components by mass spectrometry

The CMR-containing fractions collected after the last anion exchange purification step were combined and analysed by in-solution tryptic digest followed by LC-MS/MS mass spectrometry in house. This method would enable us to confirm the presence of all the CMR subunits and identify contaminants and interacting proteins.

Samples from two different CMR preparations were analysed, and from different parts of the broad MonoQ peak (see figure 3.19). Also, the presence of the SsoRNA polymerase complex in the second MonoQ peak was confirmed.

From the analysis results and the mass spec scores we can confirm that the dominant proteins in all samples are the subunits of both the CMR clusters in *Sulfolobus solfataricus*. In table 3.2 are depicted only the protein “hits” with score above 100, and only CMR components with lower scores. The two clusters reside in different parts within the genome, and are adjacent to different CRISPR loci. The first cluster (family D) consists of genes sso1424-sso1432 and is located between near CRISPR loci B and C, while the second cluster (family B) is comprised of genes sso1986-sso1992 and is neighboring CRISPR locus F. The fact that they co-purify is not unexpected since their composition indicates similar size and physicochemical characteristics. It is also possible that both complexes have bound CRISPR RNA transcript, and therefore are physically linked through their RNA load. Different sections of the broad CMR elution peak from the MonoQ column during the second CMR preparation were analysed to detect differences in the protein content. In the first part of the peak (section A in figure 3.19) both complexes seem to elute but only CMR family B is present in the late peak fractions (table 3.2). Whether this divergence reflects a difference in physicochemical properties such as weaker surface charge, or is a matter of *in vivo* abundance is unknown.

The samples contained relatively few contaminants, with most of them common in both preparations (table 3.2). The major contaminants with scores as high as Cmr2 were an AAA-ATPase (Sso0176) and reverse gyrase whose presence can be explained by their affinity for nucleic acids and their potential interaction with RNA CRISPR sequences bound to the CMR complexes. Other contaminants with a much lower score include heat shock protein 20 (Sso2427), the thermosome subunits (Sso0862, Sso0282), mRNA and rRNA processing enzymes (Sso0939, Sso0940), a second AAA-ATPase (Sso0421), ALBA and a putative DNA helicase (Sso2450). The first two proteins belong to the chaperones group (hsp20, thermosome) and would potentially be needed to assist the complex assembly. The rest are either RNA processing enzymes or known nucleic acid interacting proteins (ALBA, AAA-ATPase, potential helicase) therefore they might be interacting with any complex-bound RNA.

Subsequent protein preparations by Paul Talbot replicated these findings, and further optimisation of the purification method resulted in successful separation of the two CMR complexes.

Protein	Family	Purif A MonoQ (Mascot score)	Purif B early peak MonoQ (Mascot score)	Purif B late peak MonoQ (Mascot score)
Sso0176	AAA ATPase		1251	333
Sso1991	Cmr2	1287	1209	1061
Sso1989	Cmr1	938	654	531
Sso0420	Rev. Gyrase	653	87	
Sso1988	Cmr5	624	168	140
Sso1429	Part of Cmr2	390	256	
Sso1432	RAMP	378	262	
Sso1990	Cmr6	338	346	338
Sso1987	Cmr4	336	73	18
Sso1992	Cmr3	312	522	390
Sso1427	RAMP	287	217	
Sso1424	Cmr7?	282	256	
Sso1426	Cmr4?	268	463	
Sso1425	Cmr4?	228	110	
Sso1986	Cmr7	217		
Sso1428	Part of Cmr2	178	34	
Sso1431	Cmr3	145	25	
Sso2427	hsp20	103	221	360
Sso1430	RAMP	69		
Sso0862	Thermosome alpha subunit		344	509
Sso0939	Pre mRNA splicing protein		196	322
Sso0804	hypothetical		175	143
Sso1442	hypothetical		150	
Sso0421	AAA ATPase		132	
Sso0282	Thermosome beta subunit		120	507
Sso2450	Putative DNAhelicase		115	269
ALBA			277	259
Sso0940	Pre rRNA processing			218

Table 3.2: Mass Spectrometry analysis of two different purifications of SsoCMR.

Top hits with Mascot score above 100 are mentioned, and CMR-associated proteins if their score is lower.

3.9 Initial functional characterisation of the native SsoCMR complex

During the time we isolated the native SsoCMR complex, little or no information was available regarding the functional details of different CRISPR systems, either bacterial or archaeal. The identification of *E. coli* CASCADE (Brouns *et al.* 2008) and the *Pfu*CMR complex (Hale *et al.* 2009) had not been published at the time, therefore our initial characterization plan had to take in account all the possible roles of the complex within the CRISPR system. For this reason, it is obvious now why many of the experiments described in this paragraph had negative results, as they were designed to test hypotheses that were later proven to be incorrect. Nonetheless, these experiments are essential in the initial stages of characterisation of any hypothetical protein.

Since very limited information could be extracted from the bioinformatics analysis of the CMR complex components and the structure of Cmr7, assumptions about the possible *in vivo* function of the CMR complex could be drawn only by considering its potential role in the context of the CRISPR operation system, specifically the interference stage. To participate in this stage, the Cmr complex would be predicted to exhibit a combination of the following activities:

- affinity for nucleic acids (either RNA or DNA, CRISPR related or unspecific sequences)
- nuclease activity, in terms of attacking the extrachromosomal invader nucleic acid, or processing the CRISPR transcript (information of the processing role of Cas6 had not yet been published)
- potential nucleotide incorporation or reverse transcription activity, based on the fact that Cmr2 was predicted to be a putative novel polymerase (Makarova *et al.* 2006), as explained in paragraph 3.1.

The above possible activities were investigated with the substrate library available, which consisted of DNA and RNA versions of the repeat sequence of CRISPR locus F (CRISPR_compF), the repeat complement of locus B (CRISPR_compB), and the *in vitro* transcribed first two repeat-spacer units of CRISPR locus A (figure 3.22; clusters A and B have the same repeat sequence). In retrospect, more appropriate substrates would have been DNA or RNA protospacer sequences in order to mimic the *in vivo* substrates during the interference stage.

3.9.1 The SsoCMR complex does not bind ssDNA or ssRNA

The electrophoretic mobility shift was used to determine the ability of the native SsoCMR complex to bind nucleic acids. The limitation of these assays was that the low yield and low concentration of the complex purified by the antibody-assisted method described in the previous paragraph might not result in a strong enough

interaction to produce an observable shift. Also, we were unable to determine an accurate protein concentration since the stoichiometry of the complex is unknown and the yield was too low to perform a Bradford assay. The crude estimated concentration of the purified SsoCMR complex used in the assays described in figure 3.21 was 110 µg/ml.

The effects of temperature, concentration and pH/salt conditions were investigated. In the reactions, 1 µl of the concentrated SsoCMR pool was incubated with 1 µM of fluorescein-labelled substrate either for 20 min at room temperature/ 37°C, or 5 min at 50°C / 70°C, in binding buffer A (20 mM MES pH 6, 100 mM NaCl, 1 mM EDTA, 0.1 mg/ml BSA) or B (50 mM Tris-HCl pH 8, 100 mM KCl, 5 mM MgCl₂, 1 mM DTT). The substrates were RNA or DNA versions of the repeat complement sequence of CRISPR cluster F (CRISPR_compF), and the repeat complement of CRISPR cluster B (CRISPR_compB). No interaction was observed with either the DNA or RNA substrates under the reaction conditions tested (figure 3.21). It has to be noted that these assays did not constitute an exhaustive screening of all the possible reaction conditions although the results were repeatable. For example, a full metal and buffer screen was not carried out, as the purified protein stock was limited. It is a strong possibility that the concentration of the protein stock was not high enough to result in a stable complex with the nucleic acid substrates. However, other catalytic assays typically require smaller amounts of protein, so it was considered more useful to investigate other possible activities.

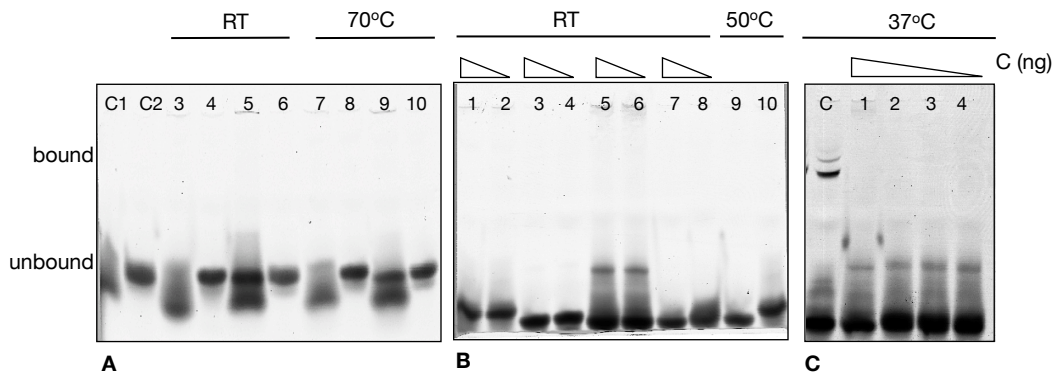


Figure 3.21: Binding of native SsoCMR to CRISPR ssDNA and RNA substrates.

Assays in gels A and B were carried out in binding buffer A. Description of gel lanes:

- (A) C1, control ssRNA CRISPR_compF; C2, control ssDNA CRISPR_compF; 3, assay with ssRNA CRISPR_compF; 4, assay with ssDNA CRISPR_compF; 5, assay with ssRNA CRISPR_compB; 6, assay with ssDNA CRISPR_compB; lanes 7-10 the same as 3-6 at 70°C.
- (B) Two-fold serial dilutions (approximately 55 ng and 27.5 ng according to our crude estimate) of the original ssoCMR sample assayed in this gel. Lanes 1+2, ssRNA CRISPR_compF; 3+4, ssDNA CRISPR_compF; 5+6, ssRNA CRISPR_compB; 7+8, ssDNA CRISPR_compB; 9, ssDNA CRISPR_compF; 10, ssRNA CRISPR_compF.
- (C) In this assay, serial dilutions of SsoCMR (55 ng, 27.5 ng, 13.7 ng and 6.8 ng) were incubated with ssRNA CRISPR_compB in binding buffer B. Lane C, positive control reaction with purified SsoCas2 which binds to ssRNA. The smeared appearance of the unbound RNA substrate in lanes 5,6 (B) and in gel (C) was most probably due to sequence

forming secondary structures and can also be seen in the control lane C1 (A), therefore they were not regarded as a real shift due to protein binding.

3.9.2 The SsoCMR complex does not exhibit nuclease activity against CRISPR RNA

In order to mediate the processing, interference or adaptation stage of the CRISPR activity, the CMR complex would need to display nuclease activity against either the CRISPR transcript or repeat sequence, or against protospacer-like sequences. Various concentrations of SsoCMR (initial estimated concentration 110 $\mu\text{g/ml}$) were incubated with radiolabelled CRISPR transcript (constructed by *in vitro* transcription of the first two repeat-spacer units of CRISPR locus A as described in Materials and Methods) for 1 hour at temperatures ranging from 37 to 65°C in 50 mM Tris-HCl pH 8.5 or 20 mM MES pH 6, 100 mM KCl, 5 mM MgCl₂, 5 mM MnCl₂, 1 mM DTT and run on a denaturing 20% acrylamide / 7M urea gel as described in Materials and Methods. The substrate was heated at 65°C for 5 min before adding the reaction mixture in order to denature secondary structures. No nuclease activity was observed under these conditions (figure 3.23, A and B). As a possible explanation it was considered that the secondary structure of the substrate could be important for the protein activity, but this result was repeatable. The protein complex was also incubated at various temperatures with a ssRNA substrate representing the complement of the repeat sequence of loci A and B (CRISPR_revB), but no activity was observed (figure 3.23 C). An illustration of the substrates used can be seen in figure 3.22. It was not considered relevant to measure nuclease activity against DNA substrates, as this would not be relevant in the CRISPR operation model.

These experiments, although they did not represent an exhaustive investigation of the potential nuclease activity of the CMR complex, indicated that it was not involved in the processing of the precursor CRISPR transcript or even the final stages of the formation of the final crRNA sequence that would mediate the interference.

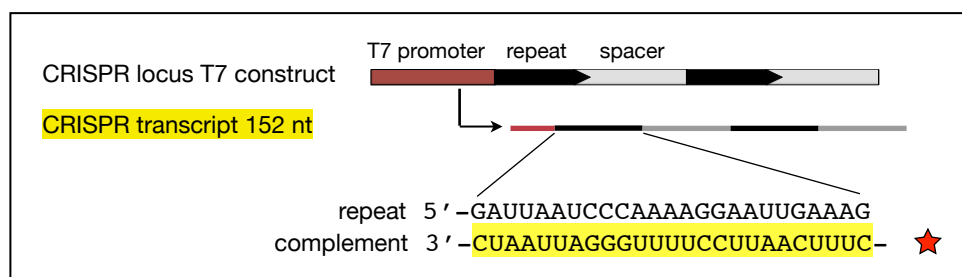


Figure 3.22: Substrates used for assaying the nuclease activity of the SsoCmr complex
A T7 promoter was cloned upstream of the first two repeat-spacer units of CRISPR locus A, in order to generate the 152 nt CRISPR transcript by *in vitro* transcription with the T7 RNA polymerase. The repeat sequence of the locus (and of locus B) is illustrated at the bottom. The substrates used in assays are highlighted in yellow. The 5' fluorescein label is indicated with a red star.

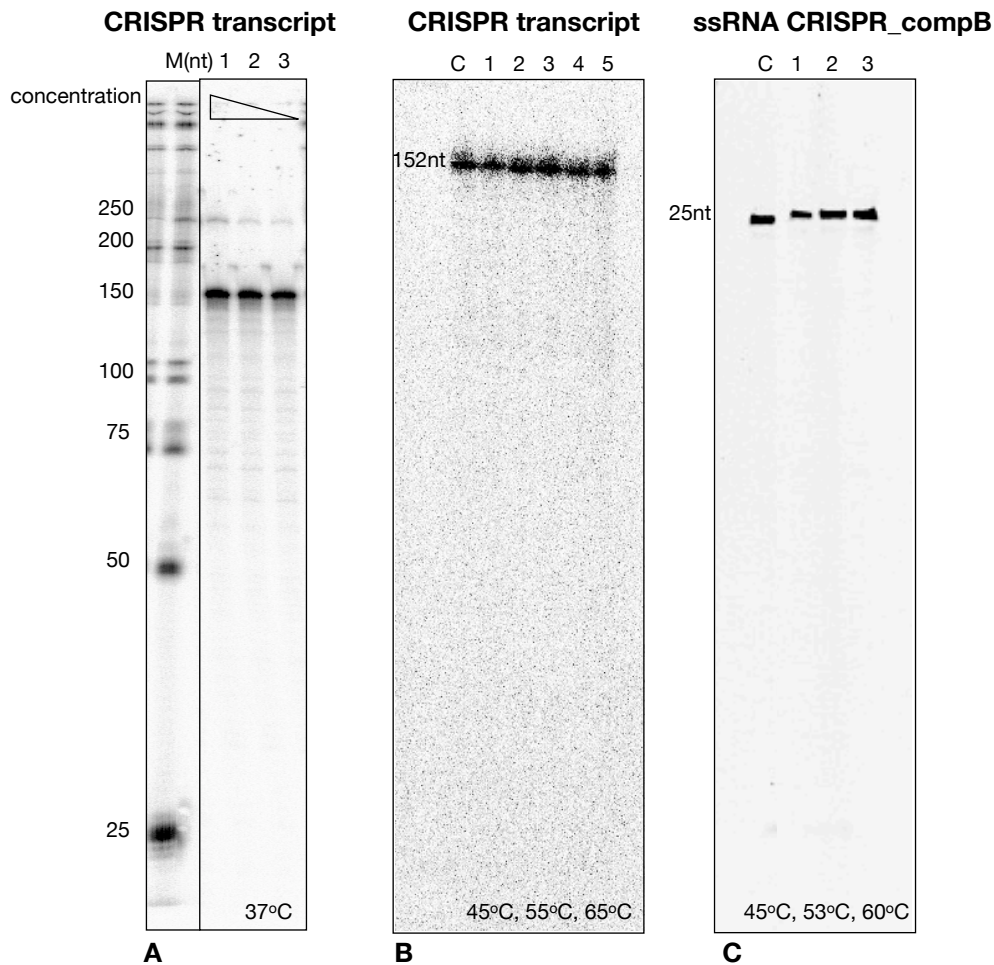


Figure 3.23: Nuclease assays of SsoCMR

(A) CRISPR transcript with two sets of repeat/spacer units (152 nt) incubated with 110, 55 and 11 ng of protein (lanes 1, 2, 3 respectively) for 1 hr at 37°C in the presence of MgCl₂ / MnCl₂. M, nucleotide size ladder. No cleavage products observed.

(B) CRISPR transcript (152 nt) incubated with 0.4 µg of protein in the presence of MgCl₂ / MnCl₂ for 30 min at 45, 55 and 65°C. C, control reaction without protein; lanes 1-3 assay at 45, 55, 65°C respectively; lanes 4-5, assay at 55, 65°C with a different protein batch. No cleavage products observed at elevated temperatures.

(C) 25 nM of fluorescein-labelled ssRNA CRISPR_compB were incubated with 0.4 µg of protein at 45, 53 and 60°C for 30 min (lanes 1-3). Lane C, control reaction without protein. No activity was observed.

3.9.3 The SsoCMR complex does not exhibit polymerase activity

The stability and size of the CMR complex, its continuous presence in the cell in amounts comparable to RNA polymerase and the fact that its largest protein component (Cmr2) contains a palm-domain polymerase motif (Makarova *et al.* 2002) initially led to the hypothesis that it might take part in the adaptation stage of the CRISPR pathway, whereby novel invader protospacer sequences are inserted into the genome immediately downstream of the leader sequence of a given locus and become part of the CRISPR spacer library. This would require some form of reverse

transcription in the case of RNA viral sequences and incorporation of the resulting DNA sequences, or the processed DNA fragments of DNA extrachromosomal elements, to the CRISPR locus.

Reverse transcriptase activity was measured by performing *in vitro* transcription with the SsoRNA polymerase as described in Materials and Methods (figure 3.24), but replacing the reverse transcriptase with the CMR complex in the primer extension step. The CRISPR locus constructs I and II cloned into the pCR2.1 TOPO vector were used as substrates for transcription, as described in Materials and Methods. These constructs contain 245 nt (CRISPR I) or 165 nt (CRISPR II) of leader sequence of the CRISPR locus and four repeat-spacer units. A successful run-off transcription reaction would result in a transcript of defined length (since the substrates were linearized), which would then be detected by reverse transcription with a radiolabelled primer, complementary to an internal site or the end of the transcript. One μl (110 ng) of the CMR complex was used in each primer extension reaction which were incubated at 70°C instead of 42°C to imitate physiological *in vivo* conditions. The reverse transcription products were separated on a 10% polyacrylamide, 7 M urea gel (figure 3.25). As can be seen in the figure 3.25 A, the sequences produced by the SsoRNA polymerase/reverse transcriptase are close to the transcript size of 274 nt (construct with four repeat-spacer units), but no products were observed with SsoCMR. The reaction was repeated over a temperature range of 50°C to 70°C with negative results (figure 3.25 B). All the appropriate controls were carried out in every assay.

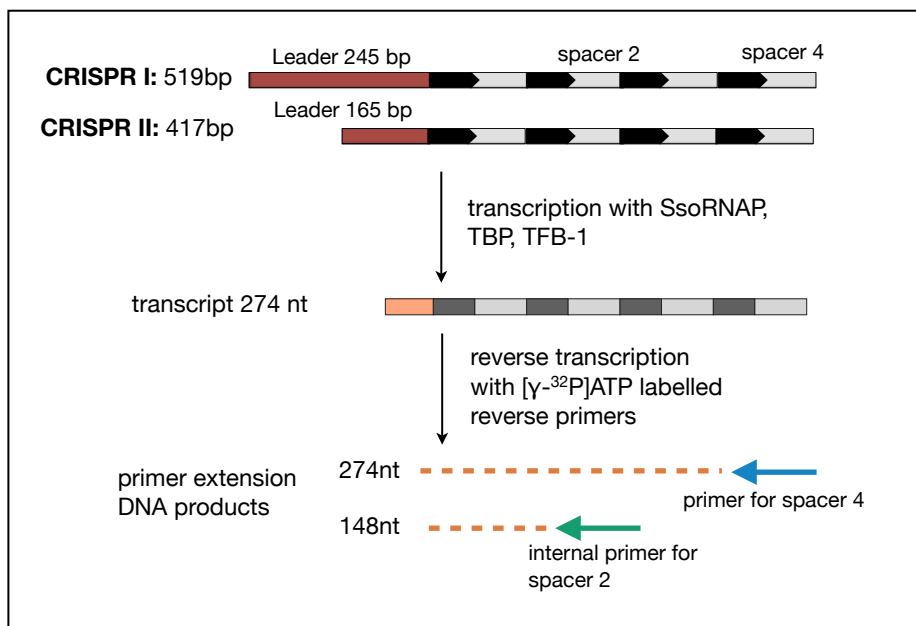


Figure 3.24: Outline of *in vitro* transcription with the SsoRNA polymerase

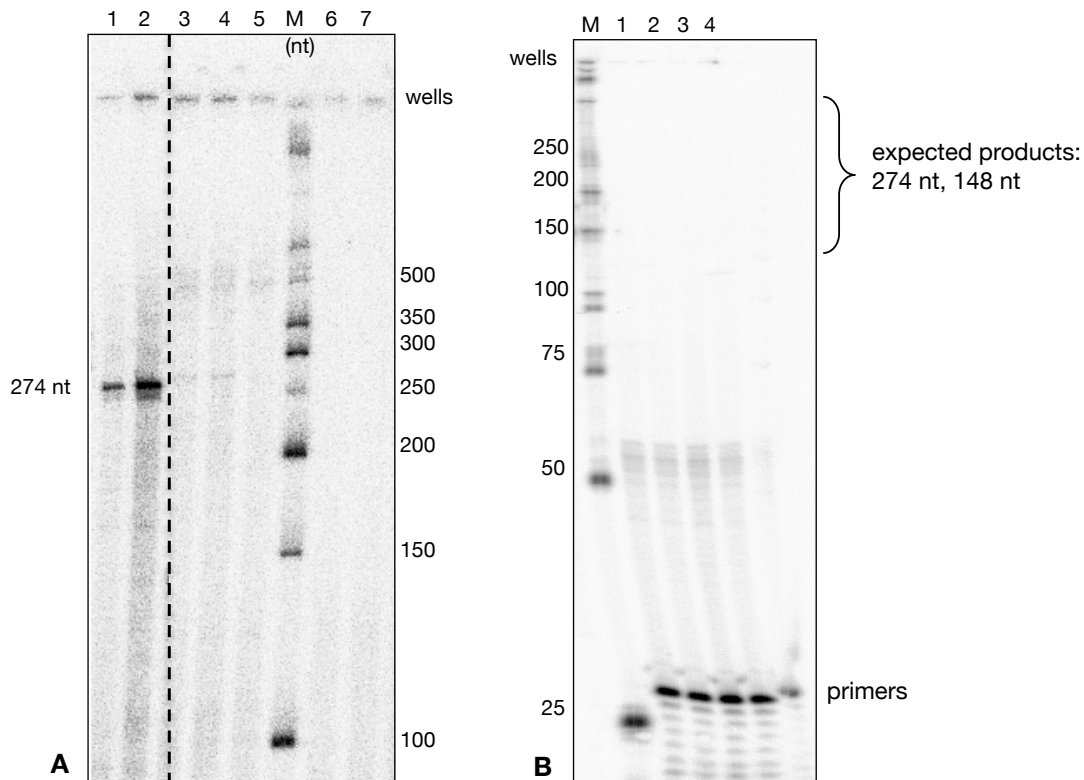


Figure 3.25: Reverse transcriptase assays for SsoCMR

- (A) *S. solfataricus* *in vitro* transcription was carried out on CRISPR locus constructs, with the primer extension fragments produced by reverse transcriptase visible in lanes 1 and 2: 1, transcription from CRISPR I, primer extension with +252r (fourth spacer); 2, transcription from CRISPR II, primer extension with +252r, complete length of transcript is 274 nt. Lanes 3, 4, 5 contain negative control reactions without TBP, TFB1 and TBP/TFB1 respectively. Reactions with SsoCMR instead of reverse transcriptase can be seen in lanes 6, 7 with CRISPR I and II transcripts as substrates respectively. M, New England Biolabs low MW DNA ladder. Dashed line indicates non-contiguous lanes on the gel.
- (B) Primer extension reactions with SsoCMR were repeated with CRISPR II transcripts at 50°C, 60°C, 70°C (lanes 1-3, respectively), and CRISPR I at 70°C (lane 4). No product was observed. The excess of radiolabelled primers can be seen at the bottom of the gel.

3.8 Discussion

The function of the CMR complex was revealed finally in an influential study by Hale *et al.* in 2009, where it was shown that the CMR complex is the effector RNA targeting agent in *Pyrococcus furiosus*. The study presented in this chapter took place before such information was published, and the experimental strategy was designed in the light of the few experimental and bioinformatical studies available until then, namely the comparative genomic analyses by Jansen *et al.* (2002), Makarova *et al.* (2002 and 2006), Haft *et al.* (2005) and the characterisation of Cas6 from *Pyrococcus furiosus* (Carte *et al.* 2008).

Hale *et al.* (2008 & 2009) demonstrated that the six Cmr module proteins from *Pyrococcus furiosus* (Cmr1-6, type III system proteins with the current classification)

co-purify with two species of mature psiRNA sequences (and few contaminating proteins), and this RNP complex is sufficient to catalyze the homology-dependent cleavage of target RNA. The mature psiRNA species are generated from the Cas6-assisted processing of the initial CRISPR loci transcripts, and consist of 37 or 31 nucleotides of spacer sequence and 8 nucleotides of the upstream repeat sequence (referred to as “psi-tag”). After cleavage by Cas6 which generates the 5' end of the psiRNA, further processing of the remaining repeat at the 3' end is taking place by an unknown protein to produce the final mature psiRNA. The Cmr-bound psiRNA is responsible for the recognition and binding of the complementary foreign RNA element, and acts as a molecular “ruler” to guide the cleavage of the RNA invader by the Cmr complex (figure 3.26). Cleavage of the target RNA occurs 14 nucleotides from the 3' end of the psiRNA regardless of the spacer length in the psiRNA species present, yields products with 3'-phosphate and 5'-hydroxy termini and is divalent cation-dependent. After reconstitution analysis, the authors concluded that all subunits with the exception of Cmr5 were necessary for target binding and silencing. The catalytic subunit performing the cleavage is not yet identified, although Cmr2/Cas10 containing an HD-nuclease domain is the strongest candidate. Thus the mechanistic details of psiRNA-guided RNA targeting in type III CRISPR systems were partially elucidated.

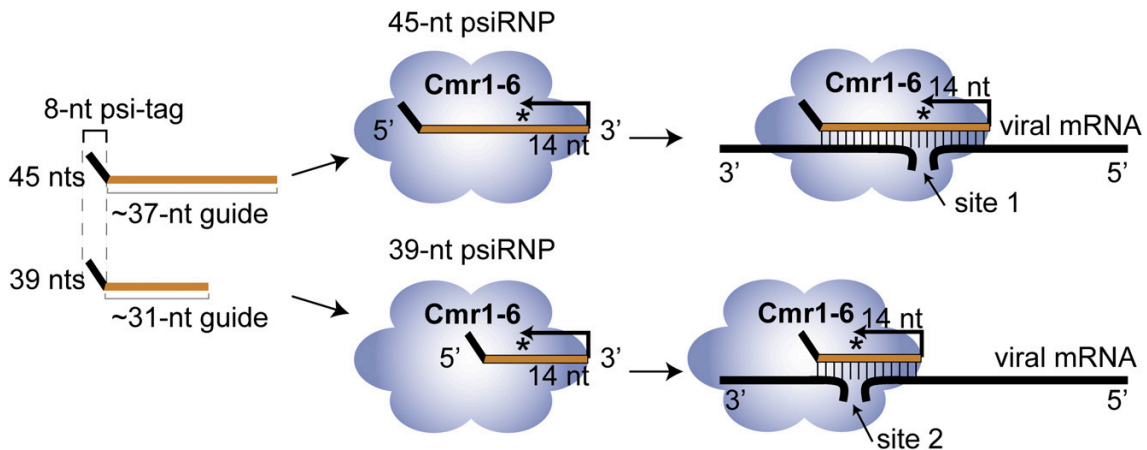


Figure 3.26: Mode of action of the PfuCmr complex.

The Cmr complex (Cmr1-6, indicated in blue) loaded with CRISPR psiRNA sequences (orange) identifies the complementary invader ssRNA (black) leading to site-specific cleavage. The psiRNA acts as a molecular ruler, positioning the target ssRNA close to the active site of the complex so that the cleavage occurs 14 nucleotides upstream from the 3' end of the psiRNA. In that way, different species of psiRNA with different lengths will cause cleavage of the target RNA at two distinct sites. Adapted from Hale *et al.* 2009.

The isolation of the native Cmr complex from *Sulfolobus solfataricus* and the initial attempts to characterize its recombinant subunits are described in this chapter. All six *S. solfataricus* *cmr* genes were cloned individually in appropriate vectors to allow for heterologous protein expression, but only Cmr1, Cmr3, Cmr4 and Cmr7 were

expressed and purified successfully during the course of this study. The remaining subunits were either insoluble (Cmr5) or did not express at all (Cmr2, Cmr6) in the heterologous *E. coli* system, indicating problems with protein stability and the lack of vital interacting protein partners. Co-expression of the unstable proteins in pairs with their adjacent *cmr* gene (e.g. *sso1991-sso1992* or *sso1990-sso1991*) using appropriate vectors such as pETDuet and pACYCDuet was also attempted with unsuccessful results. The next experimental step would be to attempt recombinant expression of the complex subunits, individually or in pairs/clusters of 3, in *Sulfolobus solfataricus* itself, employing the genetic manipulation methods made available by the work of Sonja-Verena Albers and colleagues (Albers *et al.* 2006). This method has proven successful with a number of otherwise unstable recombinant proteins or improved the catalytic activity of archaeal recombinant enzymes, and would theoretically provide the appropriate *in vivo* environment and post-translational modification machinery necessary to render the recombinant proteins stable and catalytically active. In order to achieve a complete biochemical, functional and ultimately structural characterisation of the Cmr complex, *in vitro* reconstitution from its recombinant subunits is a priority.

The stability and increased expression levels of recombinant Cmr7 were unique among the other recombinant Cmr proteins in this study. This characteristic proved extremely valuable for the isolation and study of the native Cmr complex, enabling also the crystallographic study of Cmr7. The abundance and stability of Cmr7 is explained partially by the complex stoichiometry, where we can clearly observe the excess of Cmr7 over the rest of the subunits. This fact along with its catalytic inactivity could indicate that the protein has a primarily structural role in the complex, providing a scaffold upon which the remaining subunits are assembled. Further insight into the complex stoichiometry and topology can be acquired by native mass spectrometry, single-particle EM and small-angle X-ray scattering (SAXS) analysis.

Regarding the investigation of the recombinant protein interactions, only Cmr3 and Cmr1 were clearly shown to interact. Since all of the Cmr proteins have been shown to co-purify, it can be safely assumed that they interact physically. However, the topology of the native complex can limit the reconstruction of these interactions with the recombinant proteins *in vitro*. These proteins may form part of a large protein complex but the order in which it is assembled and the individual pair interactions are still unknown. Moreover, it is reasonable to assume that Cmr2 is essential for the complex assembly, since it is the largest and also the predicted catalytic subunit. In retrospect, the CRISPR transcript or mature crRNA sequences processed by SsoCas6 (see chapter 4) would have been a more suitable candidate to mediate Cmr subunit interactions *in vitro*, instead of the CRISPR DNA repeat sequences used here.

Cmr1 demonstrated the ability to bind ssRNA, exhibiting a slight preference for the ssRNA CRISPR repeat. The interaction would be characterised as weak, although the highest concentration of Cmr1 we were able to use was 18.25 μ M, which was not high enough to give us an overview of the binding curve or a rough K_D . The substrate in retrospect was not physiologically relevant, as in the context of the Cmr complex and drawing from the PfuCmr analysis the complex would be expected to bind the mature psiRNA comprising of the 8nt repeat psi-tag and the spacer sequence. Nevertheless, it is safe to assume that the affinity of the protein for ssRNA is a result of its physicochemical characteristics and the signature ferredoxin-like fold of the RAMP superfamily it belongs to. As a result, we can speculate that the role of Cmr1 within the SsoCmr complex involves the binding and potentially recognition of the mature psiRNA and perhaps its target.

As a general comment, the weak binding properties determined by the EMSA and the failure to detect any other protein-nucleic acid interactions can be attributed to a number of factors such as unfavourable assay conditions (temperature, buffer composition, metal ion requirements), or the fact that the nature of the interaction is such that is easily disrupted.

The native SsoCmr protein complex was successfully isolated and purified from *S. solfataricus* cell lysate by four chromatographic purification steps, with antibodies raised against recombinant Cmr7 used to track the protein complex. This was made possible due to the integral stability of the complex, comparable to the SsoRNAP complex with which it co-purified for 3 of the 4 purification steps. It is almost certain that the complex co-purifies with CRISPR psiRNA species, although this aspect was not tested in this study. Further experiments including RNA isolation and sequencing would confirm the mature psiRNA sequences found in *S. solfataricus*. Mass spectrometry analysis of the purified complex in the two last purification steps (cation and anion exchange chromatography) revealed the presence of both CMR clusters in *S. solfataricus* suggesting that they are both active and are regularly expressed in the cell, indicative of a system operating in “surveillance mode”. It is unknown whether the two complexes identify and interact with a different CRISPR locus thereby supplementing each other’s function within the SsoCRISPR system or they function redundantly. Further analysis of their respective RNA content would clarify this issue. Proteins belonging to the partial Cmr clusters in *S. solfataricus* (genes sso1514-sso1510 and sso1725-sso1730) were not detected, supporting the assumption that these clusters have degenerated and are inactive.

As mentioned already in the respective paragraph, many aspects of the functional characterisation of the SsoCmr complex were designed before the information on PfuCmr was made available and the role of the complex was elucidated (Hale *et al.* 2009). In retrospect the negative results can be interpreted and

understood as the hypotheses they were designed to test were proven incorrect. The nucleic acid binding ability of the complex was examined with the wrong substrate, as DNA and RNA versions of the CRISPR repeat sequences were used instead of mature psiRNA sequences which is the physiological substrate. It has to be mentioned though that the protein material available was limited and the Cmr concentration obtained was very low, so an extensive screening would not be possible.

The absence of any nuclease activity of the native complex against the CRISPR transcript or CRISPR repeat sequences was explained with the elucidation of the role of Cas6 in crRNA processing (Carte *et al.* 2008). Further investigation of the nuclease activity of the SsoCmr complex on the grounds of the psiRNA-guided RNA cleavage activity found by Hale *et al.* in *P. furiosus* is needed. It is speculated that the SsoCmr complex could basically assume the same role in the interference stage of the CRISPR functioning, whereby it would identify and effectively silence invader RNA sequences guided by the bound psiRNAs. The mechanistic details of this activity however might vary from what is found for PfuCmr.

The only hypothetical activity unaccounted for concerns Cmr2/Cas10. Apart from the obvious hypothetical function of the HD nuclease domain in cleaving the complementary target RNA sequences, the role of the conserved polymerase/cyclase domain remains unknown. It is now clear that the Cmr complex is not involved in the adaptation stage of the system, which the universal CRISPR-associated proteins Cas1 and Cas2 are predicted to mediate, which in turn accounts for the fact that the complex did not exhibit reverse transcriptase or non-specific nucleotide incorporation activity. Future characterisation of the complex should address this issue, and perhaps the elucidation of the Cmr2 structure would enable us to gain some insight into the role of this domain.

A pressing problem that would need to be solved before any subsequent biochemical characterisation of the SsoCmr complex is possible is the low yield of the native purification method described here. High-level recombinant expression of tagged Cmr subunits in *Sulfolobus solfataricus* using the arabinose-inducible system developed by Albers *et al.* (2006) would potentially overcome this problem and enable the purification of large amounts of the Cmr complex, as well as the ability to study its expression conditions and regulation in its native host.

Chapter 4

Purification and characterisation of Csa2-Cas5a: An archaeal CASCADE-like complex for CRISPR-mediated viral defence

4.1 Introduction

4.1.1 Biochemical and structural characterisation of the *E. coli* CASCADE

As described in chapter 1, a large multimeric complex composed of Cas type I-E proteins Cse1-5e (also known as CasA-CasE) was isolated from *E. coli* and shown to be implicated in target recognition and interference in this organism (Brouns *et al.* 2008). The complex was named CRISPR-Associated Complex for Antiviral Defence, or CASCADE. Further biochemical and structural characterisation of CASCADE by Jore *et al.* (2011) enabled the understanding of the molecular mechanism which mediates specific interference in the context of the CRISPR system. Native mass-spectrometry in combination with propanol-mediated complex dissociation revealed that the stoichiometry of the intact CASCADE complex is CasA₁B₂C₆D₁E₁-crRNA₁ and has an experimental mass of 405 kDa, but also revealed the presence of the stable sub-complexes CasB₂C₆D₁E₁-crRNA₁ and CasC₆D₁E₁-crRNA₁, indicating that CasA is loosely associated with the other subunits in the periphery of the complex. Transmission electron microscopy and small-angle X-ray scattering revealed the unusual quaternary structure of the full complex, which appears to have an asymmetric “seahorse” shape 10 x 20 nm in size, and comparison with the stable sub-complexes enabled also the elucidation of the complex topology. As can be seen in figure 4.1, CasC is arranged in a semi-circular manner comprising the backbone of the complex, with CasD, CasE and CasA being attached to the “tail” end and the two CasB subunits located at the “nose” end. From the dissociation data it is demonstrated that the crRNA is strictly associated with the CasCDE core complex, and the authors suggest that it is bound either at the “tail” end of the structure, interacting with CasE, CasD and the end of the CasC backbone, in close proximity to the DNA - binding subunit CasA, or along the CasC backbone, thereby defining the

length of the assembly. Conformational change of the complex was observed upon target DNA binding.

The complex-bound crRNA was confirmed as the product of a single cleavage event by the processing endonuclease CasE, producing 5' hydroxyl and a 2', 3'-cyclic phosphate termini. This mature crRNA unit comprises of the 8 nt repeat-derived 5' psitag, a complete spacer sequence and the remaining 21 nt of the repeat forming a hairpin on the 3' end. The 5' handle appears to be a generally conserved feature of the mature crRNAs in multiple CRISPR systems (Brouns *et al.* 2008; Hale *et al.* 2009), providing a protein binding platform for the effector complexes and perhaps indicating its important role in mediating self-nonsel self discrimination.

In electrophoretic mobility shift assays CASCADE exhibited a high affinity for ssDNA and dsDNA containing sequences complementary to the complex-bound crRNA (reported K_d values were 8 and 790 nM respectively), and minimal affinity to non-target DNA resulting from the non-specific DNA-binding ability of CasA. It could also bind target ssRNA with a lower affinity. Enzymatic and chemical footprint analysis of the CASCADE binding to ss and dsDNA demonstrated that the molecular basis of the specific target recognition is the formation of an R-loop, whereby the basepairing between the spacer in the crRNA and the protospacer in the target DNA strand leads to displacement of the non-target strand (figure 4.1, C). This process was shown to be ATP-independent, enabling the system to be constantly active in an economic and efficient way without wasting the cell resources. CASCADE alone did not catalyse degradation of the target DNA, in accordance with the initial study where the presence of Cas3 was also required to inhibit phage proliferation *in vivo* (Brouns *et al.* 2008). In the proposed model, CasA is responsible for the non-specific interaction of the CASCADE with DNA, which enables the sequence-specific scanning of ss and dsDNA species for protospacer matches. Target recognition by the CASCADE-bound crRNA induces the formation of an R-loop (in the case of a dsDNA invader), where Cas3 is recruited by an unknown mechanism and hypothetically catalyses the cleavage of the invader DNA by the HD-nuclease domain. The helicase domain fused to the HD-nuclease domain in *E. coli* Cas3 is potentially implicated in unwinding the dsDNA to facilitate the R-loop formation, or in unwinding the RNA-DNA heteroduplex to enable degradation of the target DNA and perhaps rescue the crRNA. The ability of CASCADE to recognise dsDNA as its primary target is of great physiologic significance, since most invader DNA is in a double stranded form, and therefore provides a fast and effective way to silence potential threats at their source.

CASCADE components	alternative nomenclature	stoichiometry	superfamily	TIGRfam	Function (if known)
CasA	Cse1	1	YgcL	TIGR02547	unspecific DNA binding
CasB	Cse2	2	YgcK	TIGR02548	
CasC	Cas7, Cse4	6	COG1857	TIGR01869	
CasD	Cas5e	1	COG1688	TIGR02593	
CasE	Cas6e, Cse3	1	RAMP	TIGR01907	CRISPR RNA ribonuclease

Table 4.1: Composition of the *E. coli* CASCADE

Protein names in the first column are by Brouns *et al.* (2008). Protein names by Haft *et al.* (2005) and Makarova *et al.* (2011) are presented in the second column. Superfamilies and TIGRfam models presented according to Makarova *et al.* (2011).

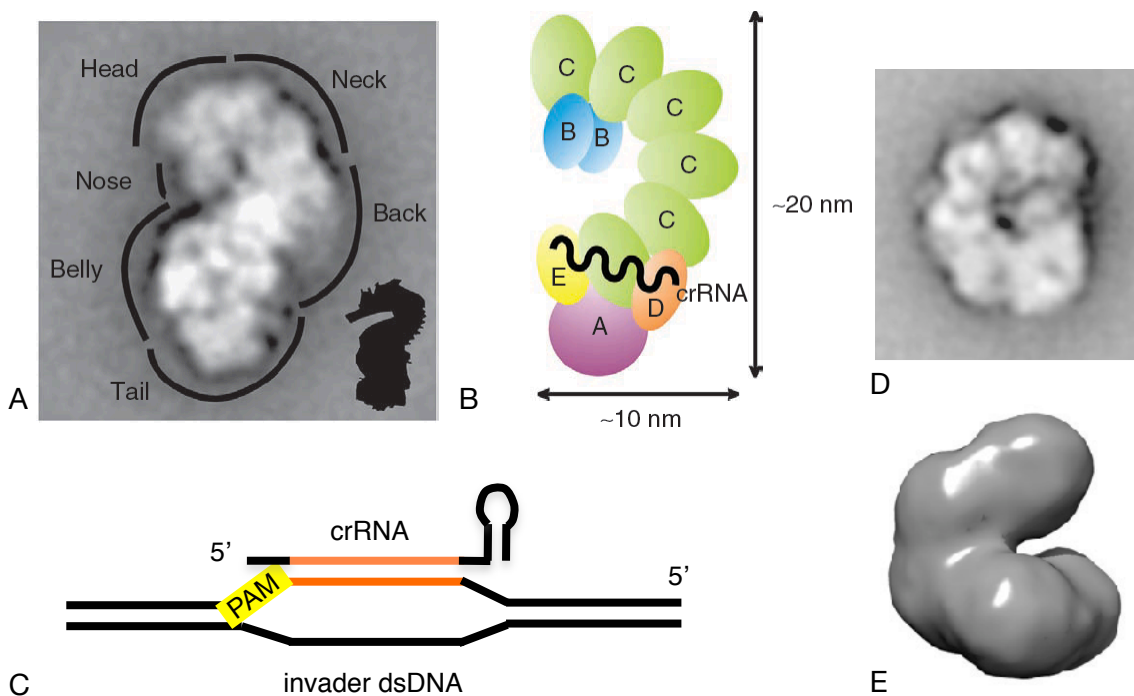


Figure 4.1: Structure of the *E. coli* CASCADE and the Csy complex from *P. aeruginosa*

(A) EM structure of CASCADE revealing the seahorse-shaped complex. (B) Structural model of CASCADE, in the same orientation as the EM image, showing the location and arrangement of the Cas subunits and the bound crRNA. Dimensions refer to the EM image. (C) The R-loop formed by CASCADE, showing the non-complementary strand of the invader DNA displaced by the crRNA, which forms a heteroduplex with the target strand. Position of the protospacer adjacent motif is shown with a yellow box, and the basepaired spacer sequence is highlighted in orange. The stem-loop secondary structure formed by the CRISPR RNA repeat of *E. coli* is shown in the 3' end of the processed crRNA. (D) EM projection and (E) SAXS reconstruction of the Csy complex, revealing the crescent-shape particle. (A)-(C) Adapted from Jore *et al.* (2011), (D)-(E) adapted from Wiedenheft *et al.* (2011).

A homologous complex was recently isolated from *Pseudomonas aeruginosa* (Wiedenheft *et al.* 2011), which harbours a type I-F system consisting of genes *cas1*, *cas3*, *csy1*, *csy2*, *csy3* and *csy4* (*cas6f*). Mass spectrometry and structural analysis by TEM and SAXS of this 350 kDa ribonucleoprotein complex revealed a subunit stoichiometry of Csy1₁:Csy2₁:Csy3₆:Csy4₁: crRNA₁, and a crescent-shaped structure 120 x 150 Å (figure 4.1 D, E). The backbone of the particle is formed by the six subunits of Csy3, and it is proposed that the crRNA molecule is bound along the arch of the complex. The main structural difference with the *E. coli* CASCADE is the lack of the “tail” observed in CASCADE where CasA is located. The result of the lack of a CasA homologue is that the Csy complex exhibits strict sequence-dependent recognition of the target and it does not have a general sequence-unspecific affinity for DNA (Wiedenheft *et al.* 2011). Similar to CASCADE, biochemical characterisation of the Csy complex also revealed crRNA-mediated target recognition within a ssDNA or a dsDNA molecule and formation of an R-loop. Moreover, the authors demonstrated that the mechanism of target recognition is based in the initial binding of a shorter seed sequence at the 3' end of the protospacer (Wiedenheft *et al.* 2011).

4.1.2 An archaeal orthologue of CASCADE

From the characterisation of the *E. coli* CASCADE described in the previous section, it becomes obvious that the minimal core of the complex consists of proteins CasC (COG1857), CasD (COG1688) and CasE, which belong to families Cas7, Cas5 and Cas6e (Makarova *et al.* 2011). Members of protein superfamilies COG1857 and COG1688 were identified very early as part of the core of the CRISPR/Cas system, since its first discovery and initial association with DNA repair (Makarova *et al.* 2002). A representative of COG1857 is present in most subtypes of what is now known as Type I CRISPR/Cas system (except I-D and I-F), encoded typically upstream of a member of COG1688. The latter was identified as Cas5, a core CAS protein by Haft *et al.* (2005), characterised by a conserved 43-amino acid N-terminal domain (TIGRFAM entry: TIGR02593) and a member of the RAM superfamily. The C-terminal part of the protein sequence shared negligible homology and was used for subtype assignment. Three conserved motifs were identified among members of the COG1857 family, namely i) s-h-N (where s and h denote small and hydrophobic residues respectively), ii) a loop containing a conserved asparagine and iii) (Phe/Pro/His/Gly)-Gly, leading Makarova *et al.* to suggest that it is an enzyme (Makarova *et al.* 2006). This conserved gene cassette (COG1857 and COG1688) was shown to be a distant homologue of *devR* and *devS* respectively from *Myxococcus xanthus* DK1622, an autoregulated gene locus involved in fruiting body development, although this observation was of

little use in predicting roles for these proteins apart from suggesting their potential physical and functional association (Makarova *et al.* 2002).

The high level of conservation across CRISPR subtypes, conserved gene synteny displayed by these two protein families and the elucidation of their key role in CASCADE has led to the establishment of Cas5 and Cas7 family proteins as core elements within CRISPR type I systems (Makarova *et al.* 2011a, 2011b).

It is unsurprising therefore that only these protein families have homologues across CRISPR/Cas type I system subtypes. In *Sulfolobus solfataricus* P2 we encounter a coexistence of types I-A (formerly known as Aperm subtype) and III-B CAS systems interspersed between six CRISPR loci, in which three paralogues of Cas5a/Cas7 are encoded (ORF numbers sso1441/sso1442, sso1400/sso1399 and sso1998/sso1997 respectively, figure 4.2). All Cas7 protein sequences contain the superfamily conserved motifs (alignment in Appendix II). In this subtype (I-A), an additional protein (Csa5, TIGR01878) is always found encoded upstream from Cas7. No functional prediction exists for this small protein (~150aa) although its genomic context indicates a potential association with the archaeal CASCADE homologues. The rest of the CRISPR gene locus (I-A) comprises the Cas3' (only the helicase domain) and Cas3" (HD-nuclease) homologues, Cas8a2 in one case, Cas6 (the predicted processing endonuclease), Csa3, Cas4, Cas2, Cas1 and Cas4, in a generally conserved order. Four type I-A cassettes are encoded in the *S. solfataricus* P2 genome (figure 4.2), but not all of them contain the full series of aforementioned genes or in some cases they are separated by non-CRISPR related genes.

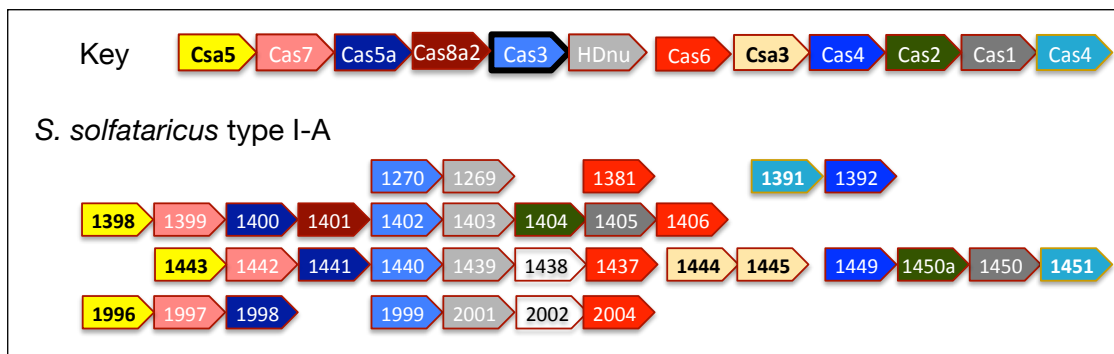


Figure 4.2: Gene names and operon organisation of type I-A CRISPR/Cas in *S. solfataricus*

In this chapter it will be demonstrated that Cas5 and Cas7, referred to as Cas5a (or SsoCas5a) and SsoCsa2 (or SsoCsa2) in *S. solfataricus* P2, form a stable protein complex *in vitro* and *in vivo* which is proposed to be the archaeal counterpart of CASCADE (aCASCADE). Initial functional characterisation of the complex components will be presented alongside structural data obtained in collaboration with Nathanael Lintner and Dr. Martin Lawrence (Montana State University Bozeman), in

order to elucidate its role in CRISPR interference and understand the molecular basis of crRNA-mediated DNA recognition in this subtype.

4.2 Site-directed mutagenesis of Csa2

Multiple sequence alignments and structure-based threading enabled the identification of conserved residues within the Cas7/Csa2 family. Conservation among Cas7 family proteins is limited, and although the three characteristic Cas7 superfamily motifs described by Makarova *et al.* (2006) are present, they do not represent suitable candidates for catalysis. Conserved residues among the Csa2 sub-family and their location on the Csa2 structure will be discussed later, but of all residues, a histidine at position 160 was selected for mutation analysis after close investigation of a DALI structural alignment between the RRM domains of SsoCsa2 and *P. furiosus* Cas6 (PfuCas6). The alignment indicated that the histidine-66 in *P. furiosus* Cas6 is equivalent structurally to *S. solfataricus* histidine-160, even though it is not part of the RNA-recognition motif which is the only conserved domain between the two proteins. The location of this residue however in regard to the RRM domain is similar, residing above the β -sheet of the RRM fold in a position potentially favourable for nucleic acid interaction (figure 4.21).

To investigate the role of the Asn cluster in nucleic acid binding, histidine-160 was mutated to an alanine by site directed mutagenesis to generate the mutant Csa2 H160A as described in Materials and Methods. A multiple sequence alignment of conserved residues among Csa2 family members can be found in Appendix II.

4.3 Expression and purification of recombinant wild-type and mutant Csa2 and the Csa2-Cas5a complex

To determine whether Csa2 and Cas5a form a stable interaction *in vivo* the genes encoding *S. solfataricus* Csa2 (sso1442) and Cas5a (sso1441) were amplified by PCR from genomic DNA and cloned into pDEST14 (for individual expression of Csa2) and pRSFDuetHISTEV (for co-expression of Csa2-Cas5a) as described in Materials and Methods. The constructs were fully sequenced to confirm their integrity. IPTG-induced protein expression was achieved in *E. coli* BL21 (DE3) host cells as described. Purification of both wild-type and mutant Csa2 consisted of a 3-step purification involving affinity chromatography on a 5 ml HisTrap HP column, size exclusion chromatography on a HiLoad 26/60 Superdex 200 and a final affinity step on a 5 ml HiTrap Heparin HP (figure 4.3, E, F, figure 4.4). The same procedure was employed for the purification of the Csa2-Cas5a complex (figure 4.3, A-D) except for an additional heat step at 65°C for 20 min before the nickel-chelate affinity

chromatography in order to precipitate the majority of the mesophilic *E. coli* proteins. The identity of the proteins was confirmed by mass spectrometry analysis. The calculated molecular weights of 38,3 kDa (plus his-tag), 35,26 kDa (native) for Csa2 and 30,44 kDa (plus his-tag) for Cas5a were confirmed on SDS PAGE.

Expression levels for both Csa2 and the complex were high, although a significant portion of the expressed protein was found in the insoluble fraction. SDS PAGE analysis revealed severe degradation products in the case of individual Csa2 expression (figure 4.3 F), which were identified as such by mass spectrometry. Measures taken to avoid protein degradation included the addition of up to 20% glycerol in all the buffers, the rapid completion of the purification procedure and the execution of all the purification steps on ice. Protein breakdown was reduced but it could not be completely avoided, presumably because the protein's stability was compromised due to the lack of interacting partner, Cas5a. It was found that the Csa2-Cas5a complex exhibited greater stability when purification procedures and short-term storage were carried out at room temperature (~25°C). Typical yields were approximately 2.6 mg/L of culture for Csa2 and 0.5 mg/L of culture for the Csa2-Cas5a complex.

Elution volumes of both the complex and Csa2 did not correspond to their theoretical molecular weight or provide accurate indications of their oligomeric state in solution, as is explained in paragraph 4.5. Furthermore, the behaviour of the proteins was inconsistent in each run, which explains the similar elution volume observed in figure 4.3 B, E.

The fact that Cas5a and Csa2 co-purify over three chromatographic columns confirmed their physical and functional association as a stable protein complex. Expression levels for Csa2 were higher than for Cas5a during co-expression, even though *csa2* was cloned in the second cloning site further away from the inducible promoter. The excess of Csa2 seemed to be separated from the complex during the third purification step through a heparin column, from which it eluted at an earlier stage, potentially as a result of differentially exposed protein surfaces.

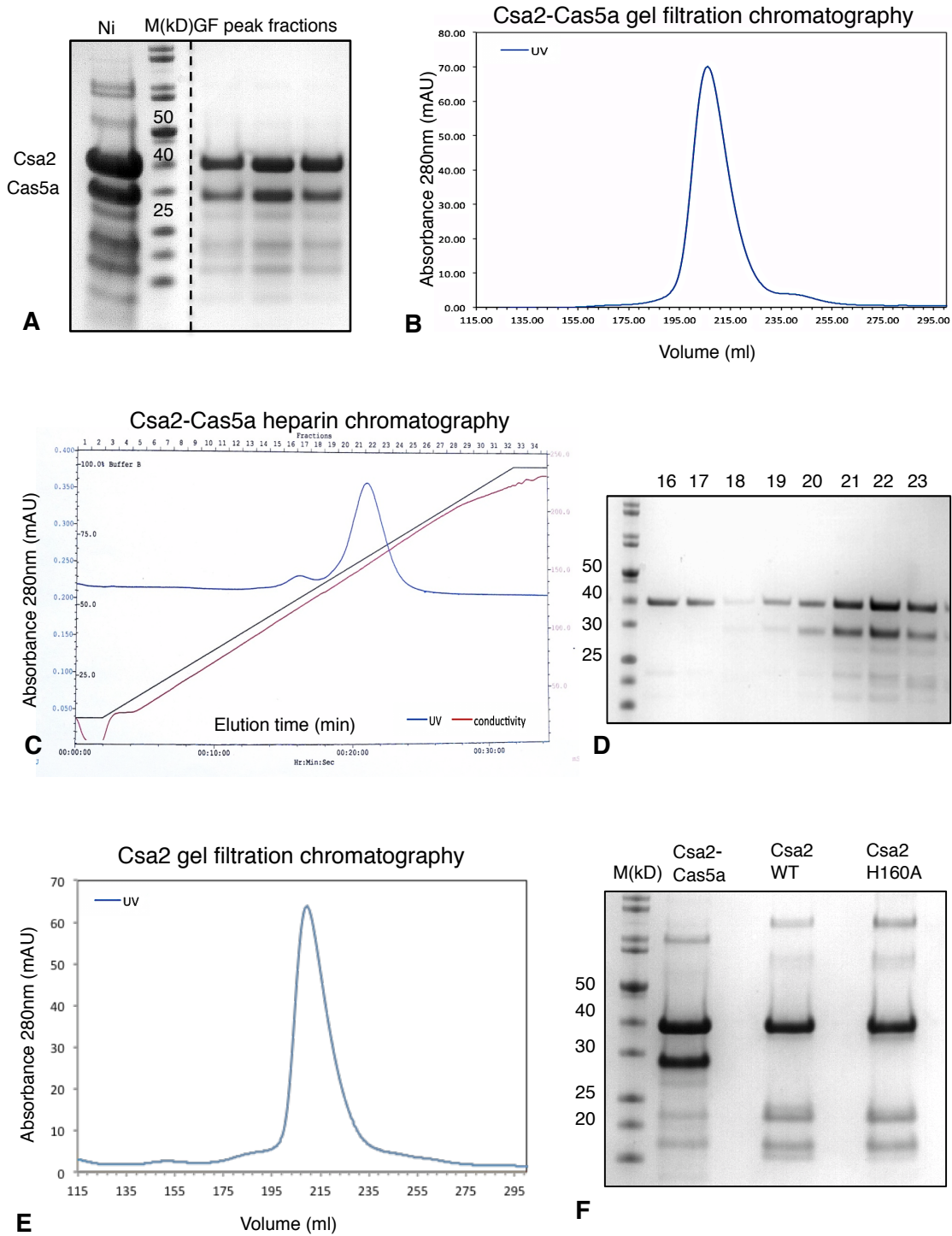


Figure 4.3 : Purification of recombinant Csa2-Cas5a and Csa2 WT, H160A.

(A) Purification stages of Csa2-Cas5a monitored by SDS-PAGE, lanes represent protein sample after nickel-chelate (Ni) and gel filtration chromatography. (B) Chromatogram of Csa2-Cas5a on a Superdex 200 gel filtration column. (C) Chromatogram of Csa2-Cas5a purified by heparin affinity chromatography. (D) Fractions of Csa2-Cas5a eluting from a heparin column, lane numbers correspond to elution fractions of (C). The excess of Csa2 is eluting as a small peak prior to the complex elution. (E) Purification of Csa2 WT by gel filtration chromatography on a Superdex 200. (F) Purified samples of the Csa2-Cas5a complex, Csa2 WT and Csa2-H160A on SDS-PAGE. Samples contain 70 pmoles of protein. The bands at ~19 kDa and ~22 kDa in the Csa2 WT/H160A samples were identified by mass spectrometry as N-terminal degradation products.

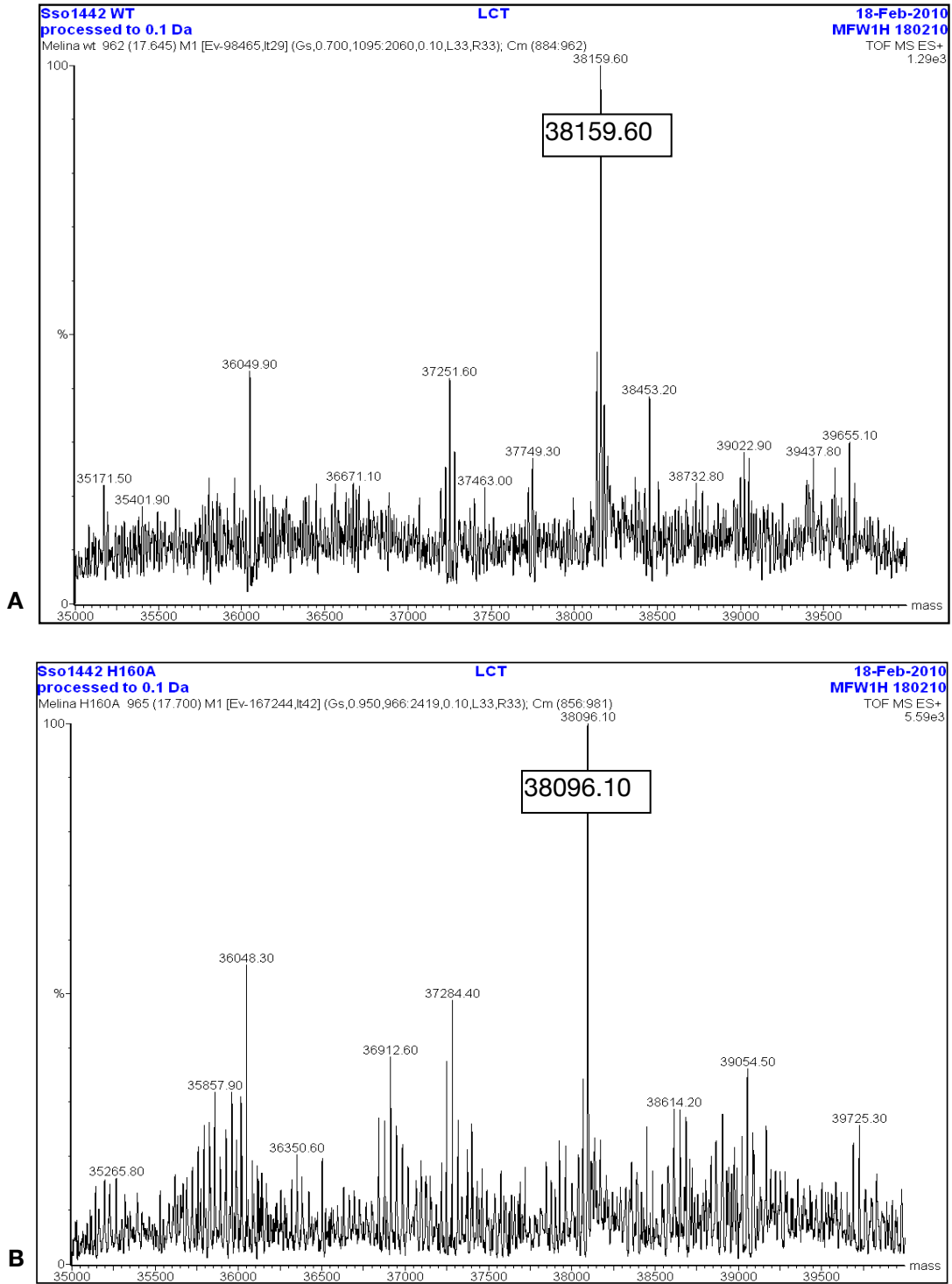


Figure 4.4: ESI-TOF mass spectrometry of Csa2 WT and H160A

Intact molecular weights of proteins determined by ESI-TOF mass spectrometry to confirm the site mutation. (A) Mass spectrum for the WT protein. The major peak corresponds to a molecular weight of 38159.6 Da (data processed to 0.1 Da). (B) Mass spectrum for Csa2-H160A. The major peak corresponds to a molecular weight of 38096.1 Da (data processed to 0.1 Da). The mass difference of 63.5 Da is in agreement with a replacement of the histidine-160 with an alanine.

4.4 Investigation of the native Csa2-Cas5a complex from *Sulfolobus solfataricus* P2 and its accessory proteins

The native SsoCsa2-Cas5a protein complex was successfully isolated by Nathanael Lintner by expressing N-terminal affinity tagged Csa2 in *S. solfataricus* strain PH1-16 under the control of the araS promoter as described in Lintner *et al.* (2011). Csa2 carried a streptavidin and 8-histidine tag which enabled a two-step affinity purification using a Strep-Tactin and Ni-NTA resin, and finally a gel-filtration chromatography step on a Superose-6 column. Visualisation of the purified sample on SDS-PAGE and mass spectrometry analysis revealed the co-purification of Cas5a with minor contaminants (figure 4.5), confirming the *in vivo* presence of a stable CASCADE-like complex, for which the term “aCASCADE” is proposed (Lintner *et al.* 2011).

To determine whether the Csa2-Cas5a complex was interacting with accessory Cas proteins that would potentially perform other aspects of a CASCADE-like function, samples of native aCASCADE partially purified only through the Strep-Tactin resin were subjected to in-solution tryptic digest and LC-MSMS. Identified proteins within the top 20 Mascot hits included the Csa2 and Cas5a paralogues, Sso1399 and Sso1440 respectively, Csa5 (Sso1443), Cas6 (Sso1437) and Csa4 (Sso1401/Cas8a2). Contaminants included biotin carboxylase (Sso2466) as it interacts with the strep resin and Alba (Sso0962), the ubiquitous nucleic acid binding protein. These findings were repeatable in three independently purified protein samples, the results of which can be seen in table 4.2. To determine whether the co-purifying proteins represented transient or more stable interactions, the protein samples were further purified by gel filtration chromatography on a Superose 12 10/300 GL (GE Healthcare). Only the core aCASCADE proteins (Sso1442/Sso1441) and their paralogues (Sso1399/Sso1400) were detected by mass spectrometry in the purified samples, indicating a weak and transient interaction with Cas6, Csa5 and Csa4.

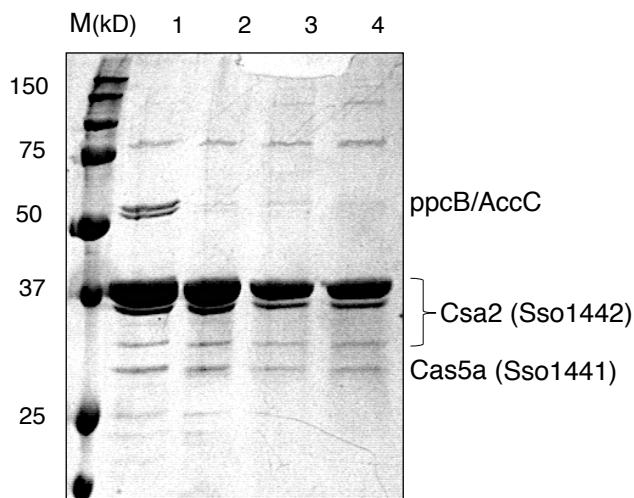


Figure 4.5: Isolation of native aCASCADE from *S. solfataricus*

(A) Stages of native complex purification on SDS-PAGE: 1, sample after StrepTactin column; contaminants such as ppB and AccC are visible above the 50 kDa protein marker band. 2, sample after Ni-NTA purification; 3 and 4, purified sample after size exclusion chromatography. The multiple bands observed for Csa2 represent the tagged and endogenous untagged form of the protein, along with possible degradation products. We can observe the strong band of Csa2 which corresponds to the excess of Csa2 over Cas5a. Adapted from Lintner *et al.* (2011).

Taking into account the role of Cas6 within the CRISPR system as the processing endonuclease of the primary CRISPR transcript, it can be suggested that the observed interaction is mediated by the presence of the CRISPR transcript in its various processing stages. This also explains the detection of Alba in the partially purified sample. Csa5 (Sso1443) is a small, basic (predicted pI 8.71), hypothetical protein 150aa in length. The gene encoding for Csa5 is located adjacent to the *csa2* and *cas5a* genes in conserved order in type I-A systems, possibly reflecting an aCASCADE accessory subunit.

Protein	Family	Expt A		Expt B		Expt C	
		Mascot score	coverage	Mascot score	coverage	Mascot score	coverage
Sso1442	Csa2	24937	78%	33795	80%	32149	82%
Sso1997	Csa2	10012	49%	17988	57%	15410	52%
Sso1441	Cas5a	2040	53%	1770	57%	1726	63%
Sso1399	Csa2	999	66%	1189	40%	912	50%
Sso2466	Biotin carboxylase	593	33%	67	6%	377	59%
Sso1400	Cas5a	396	41%	622	22%	327	22%
Sso1443	Csa5	130	25%	372	65%	377	59%
Sso0962	Alba	95	53%	173	72%	173	46%
Sso1437	Cas6	92	17%	132	17%	61	18%
Sso1401	Cas8a2	-	-	153	10%	84	8%
Sso2004	Cas6	-	-	132	15%	-	-
Sso1998	Cas5a	-	-	60	10%	90	16%
Sso1996	Csa5	-	-	-	-	77	29%

Table 4.2: Co-purifying Cas proteins in the partially purified native aCASCADE sample as identified by solution trypsin digestion followed by LC - MS/MS.

Sso1997 is a Csa2 orthologue that is 92% identical to Sso1442. Therefore this hit probably represents Sso1442 in the sample, as the sequence coverage is not sufficiently high to distinguish between the two and no unique peptides for Sso1997 were found in this sample.

4.4.1. Nucleic acid content of the native aCASCADE

The native aCASCADE purified from *S. solfataricus* was examined by Nathanael Lintner for co-purifying nucleic acid, in accordance with the crRNA and DNA content of the *E. coli* CASCADE (Brouns *et al.* 2008). An RNA species of 60-70 nt was found to co-purify with the native complex over three chromatographic purification steps, showing no visible degradation when exposed to ribonuclease treatment, thus indicating its tight association with the complex. Additional co-purifying nucleic acid species included low amounts of higher molecular weight RNA (2x or 3x the main species) and smaller amounts of ~300 nt DNA (figure 4.6, A and B).

To identify the origin of the major RNA species the 60-70nt bands were extracted, cloned and sequenced as described in Lintner *et al.* (2011). The cDNA sequences were indeed CRISPR-derived, with eight of the 16 sequenced clones featuring a complete spacer sequence with 8 nt of repeat sequence at the 5' end and 16-17 nt at the 3' end, thus representing a complete repeat-spacer unit (figure 4.6 E). These crRNA fragments are analogous to the mature crRNA species produced by the PfuCas6 and the *E. coli* casE, and could represent the mature crRNA species in *S. solfataricus*. Whether the SsoCas6 is able to generate these fragments is examined later in this chapter.

The repeat sequences in the cDNA clones represented all three of the repeat sequences encountered in the *S. solfataricus* P2 CRISPR loci, indicating that the aCASCADE is able to recognise and interact with all the active *S. solfataricus* CRISPR loci. The co-purifying DNA was not sequenced, therefore it is not known whether it represents nonspecific DNA interacting with the complex, or it is recognised by the bound crRNA in a sequence-specific manner.

Samples from the native aCASCADE provided by N.Lintner in Montana State University were also analysed for their nucleic acid content in the White lab to verify the aforementioned results. Basic phenol-chloroform extraction was performed on native StrepTactin-purified aCASCADE samples prior to size-exclusion chromatography on a Superose 12, and in two different parts of the broad elution peak (see figure 4.7 D). The extracted nucleic acid was labeled with [γ - 32 P] ATP and run on a 20% polyacrylamide / 7M urea gel. Sequencing of the fragments was not carried out due to time constraints. In figure 4.6 C, we can observe that the main component is indeed an RNA species 60-70 nt in length (confirmed by RNase A treatment), with a faint band of 130-140 nt. No changes in the RNA content of the complex were observed in the different samples, indicating firstly that this 60-70 nt species is an integral part of the complex, and secondly that it is not responsible for the puzzling elution profile of the complex on the Superose 12 column (discussed in paragraph 4.5).

The recombinant Csa2-Cas5a complex as was expected did not contain any nucleic acids, as indicated by the purified sample absorbance at 260nm.

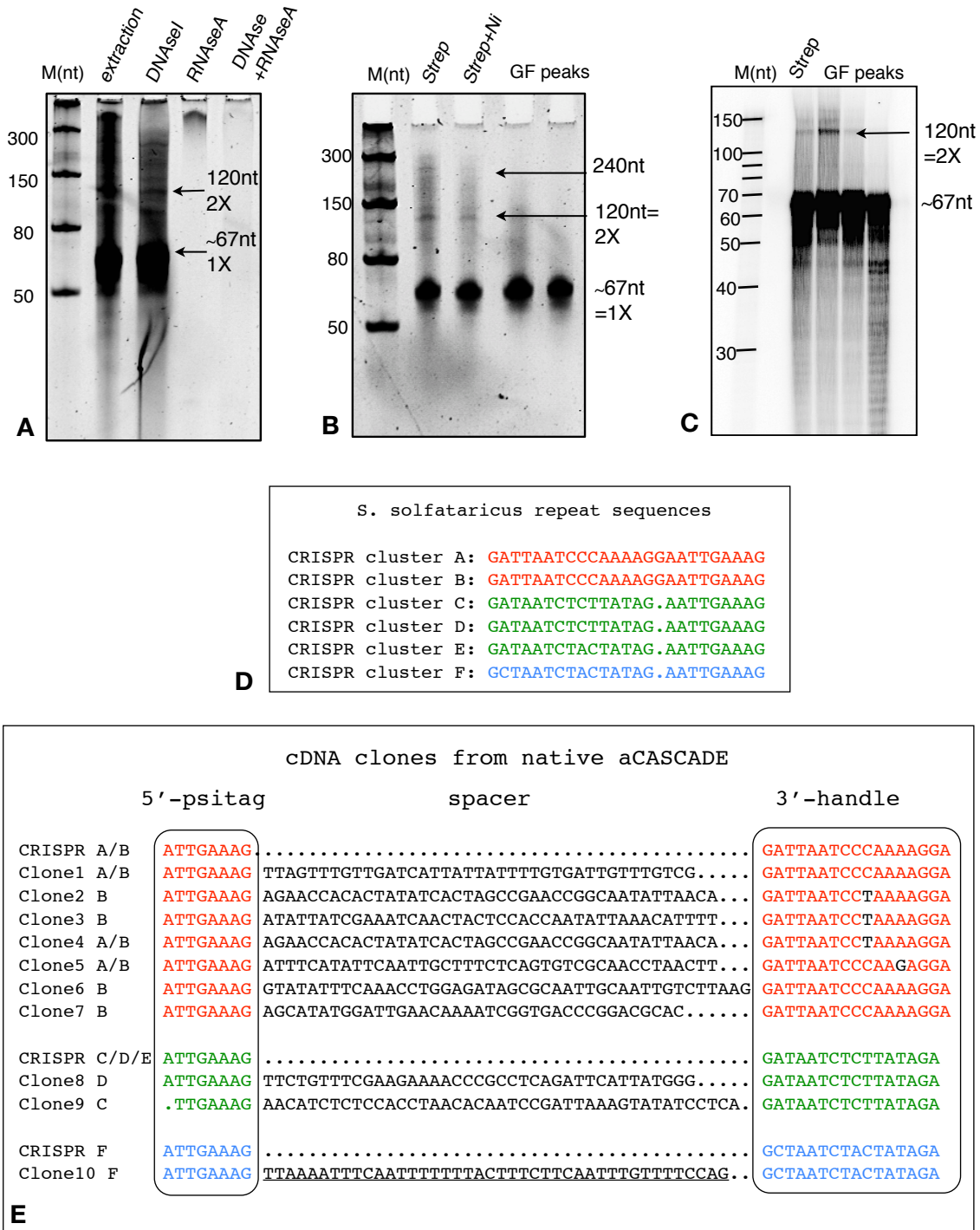


Figure 4.6: Nucleic acid content of the aCASCADE

(A) Basic phenol-chloroform extraction of nucleic acids from StrepTactin-purified native aCASCADE, ran on denaturing PAGE. Identity of the nucleic acids was verified by DNaseI and RNAseA treatment, indicated on the respective lane. The second and third lanes are overloaded to compare the relative amounts of bound DNA and RNA. The predominant RNA species are marked with an arrow. (B) Denaturing PAGE illustrating the RNA co-purification with the aCASCADE through the purification procedure. The purification sample the RNA was extracted from is indicated on the respective lane. We can observe the gradual disappearance of the 4X and 2X RNA species, due to the isolation of the core aCASCADE (Csa2-Cas5a) from Cas6 and the other accessory (and potentially RNA-binding) proteins. (C) RNA extraction from

purification samples of the native aCASCADE carried out in the White lab, ran on denaturing PAGE. The predominance of the 67 nt mature crRNA species is confirmed, along with the presence of a small amount of the 2X intermediate species. (D) Consensus repeat sequences from all active loci in *S. solfataricus*, colour-coded according to the CRISPR families by Lillestol *et al.*, (2009). Red, family II; green, family I, blue, family I cluster F. (E) Alignment of non-redundant cDNA sequences from sequencing of the predominant aCASCADE-bound RNA species. This species is revealed to be the mature processed form of CRISPR transcripts. Repeat-derived sequences in both ends of the clones are colour-coded according to (D), and suggest that aCASCADE is associated with all active CRISPR loci products. The underlined spacer matches to a protospacer in *Sulfolobus icelandicus* Rod-shaped Virus (SIRV). All figures except (C) and (D) adapted from Lintner *et al.* (2011).

4.5 Size determination and stoichiometry of the native and recombinant Csa2-Cas5a complex

In order to determine the size of the Csa2-Cas5a complex and therefore estimate the subunit stoichiometry, samples of native and recombinant complex were subjected to analytical gel filtration on a calibrated Superose 12 10/300 GL column (GE Healthcare). The column was equilibrated with 20 mM MES pH 6, 250 mM NaCl, 1 mM EDTA, 0.5 mM DTT and sample volume was 100 μ l, at a concentration of approximately 2 mg/ml. The elution volume of the complex in each sample was used to estimate its apparent molecular weight by comparison with a standard calibration curve, plotted as described in Materials and Methods.

The elution profiles of two independent native complex samples produced inconsistent results. In the first purification the complex eluted over two consecutive peaks extending from an estimated size of 571 kDa into the void volume, while during the second purification it eluted over a broad peak corresponding to an apparent mass range of 50-550 kDa (figure 4.7 D). Samples from each peak were analysed by SDS-PAGE and in-solution LC-MSMS and verified the presence of the complex. Recombinant Csa2-Cas5a complex samples that were also subjected to analytical gel filtration on the same column exhibited a different behaviour, eluting quite late from the column with apparent masses of 24 - 34 kDa, indicating possible dissociation of the complex during the run. The calculated molecular weight of the complex with a 1:1 stoichiometry should be 65.6 kDa. When recombinant his-tagged Csa2 was passed through the calibrated Superose 12 the results were again inconclusive, as the apparent masses in each run varied from monomer to dimer. Analytical size exclusion chromatography was also performed on the native Csa2-Cas5a complex expressed in *S. solfataricus* by our collaborator, Nathanael Lintner in Montana State University. The native complex eluted from a Superose 6 column (GE Healthcare) mainly in a broad peak corresponding to a molecular weight range of 350-500 kDa (Lintner *et al.* 2011).

In order to resolve this ambiguity, samples of the recombinant Csa2-Cas5a complex and Csa2 from *E. coli* expression were sent for analytical ultracentrifugation (AUC) at the University of Dundee. Results for Csa2 revealed that it exists in solution

primarily as a monomer, with relatively small amounts of dimer and trimer and a dissociation constant of 4.5 μM . For the Csa2-Cas5a complex, the major consistent peak in the sample had a mass of 54-59 kDa, which corresponds roughly to a dimer with a 1:1 stoichiometry of Csa2 to Cas5a. A range of oligomeric species was also detected, but they represented a small percentage of the sample. These results should be regarded with caution as the presence of 10% glycerol in the sample caused some interference with the measurements (data not shown).

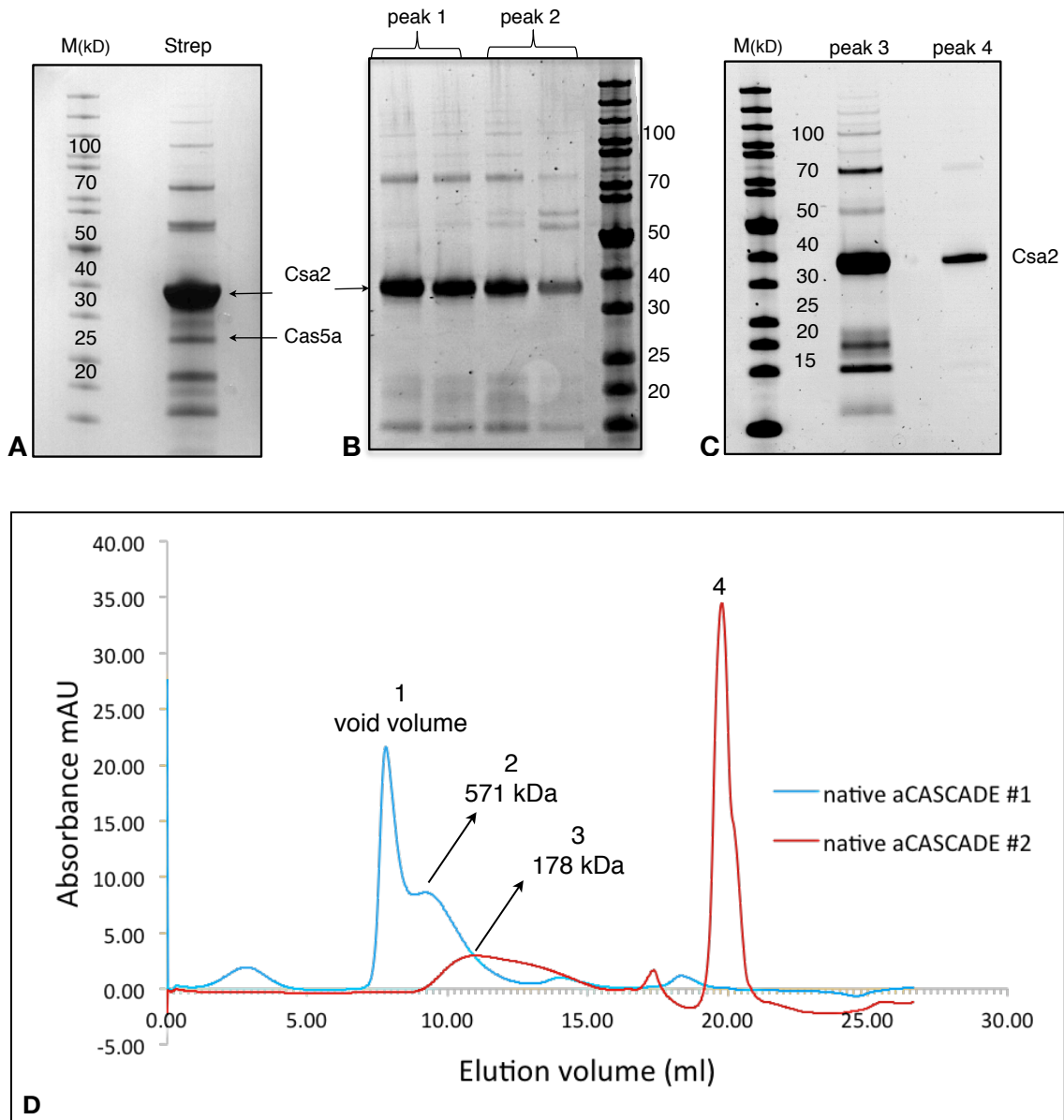


Figure 4.7: Analytical size-exclusion chromatography on native and recombinant Csa2-Cas5a complex.

(A) SDS-PAGE of the native aCASCADE sample prior to size-exclusion chromatography on Superose 12, where multiple contaminant bands are apparent. In (B) and (C), protein samples from the respective elution peaks in (D) are run on SDS-PAGE. The presence of Cas5a in the purified samples was verified by LC-MS/MS even though it is not detectable on the SDS-PAGE. (D) Superimposition of the elution profiles of two independent native aCASCADE samples on Superose 12. In the second run, the excess of Csa2 seems to have dissociated from the complex and eluted as a separate peak (peak 4).

From the above results, we are unable to draw conclusions regarding the stoichiometry of the aCASCADE although from the AUC results it seems that the predominant species in solution is a dimer. However, taking into account the RNA content of the native complex, its affinity for nucleic acids and the TEM images (discussed in subsequent paragraphs) it is possible that higher order structures can form via RNA bridging or by recruitment of extra Csa2 subunits, which would explain the broad elution peaks during analytical gel filtration and the excess of Csa2 observed on SDS PAGE (Lintner *et al.* 2011). In contrast, the recombinant complex does not co-purify with crRNAs as it would not recognise *E. coli* CRISPR sequences and therefore it should only exist as a dynamic monomer-dimer equilibrium, which is the case observed both in AUC and the late-eluting material on Superose 12.

4.6 *In vitro* protein interactions of the recombinant Csa2-Cas5a complex.

In order to investigate the protein interactions between the core Csa2-Cas5a complex and the co-purifying CAS proteins, we used paramagnetic precharged nickel particles and tagged/untagged versions of the recombinant proteins as described in Materials and Methods. The following combinations were assayed, in presence and absence of 1 µg of CRISPR RNA transcript:

- His-tagged Csa2-Cas5a with Cas3 (Sso1440) (figure 4.9 A)
- His-tagged Csa2-Cas5a with Csa5 (Sso1443) (figure 4.9 A)
- His-tagged Csa2-Cas5a with Cas3, hisCas6 (Sso1437) and Csa5 (figure 4.9 B)

12 µg of the tagged complex was bound to 20 µl of nickel particle solution in binding buffer and then incubated with 12 µg of the appropriate partner(s) for 15min at 45°C. For Cas3, 1 mM ATP/MgCl₂ was added to the reaction. After thorough washing with increasing concentrations of salt (NaCl) and imidazole, the bound proteins were eluted with 500 mM NaCl and 500 mM imidazole and visualised on SDS-PAGE. A schematic diagram of the assay can be seen in figure 4.8. Appropriate controls were carried out to ensure non-interaction of the untagged proteins and the nickel particles. No interactions were observed between the recombinant proteins either in the presence or absence of CRISPR transcript. Considering that the interactions identified during the purification of the native aCASCADE by mass spectrometry were weak and transient enough to be disrupted in subsequent purification steps, this result is not surprising. If the observed interactions *in vivo* are indeed a result of RNA bridging, then perhaps the amount of CRISPR RNA used was not sufficient to lead to a stable interaction *in vitro*, at least in the case of Cas6. It is also possible that the accessory subunit Csa5 is recruited to the complex to perform a specific function and is not constantly physically interacting with Csa2 or Cas5a. As for Cas3, it is hypothetically

recruited to the complex accompanied by the HD nuclease after the target recognition in order to degrade the invader DNA, and therefore it was not identified as an interacting partner neither *in vivo*.

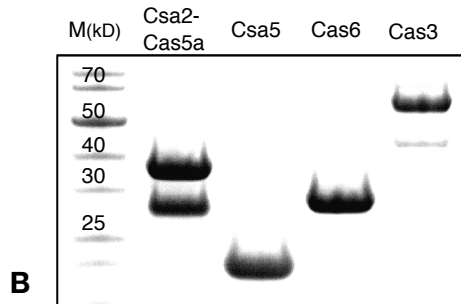
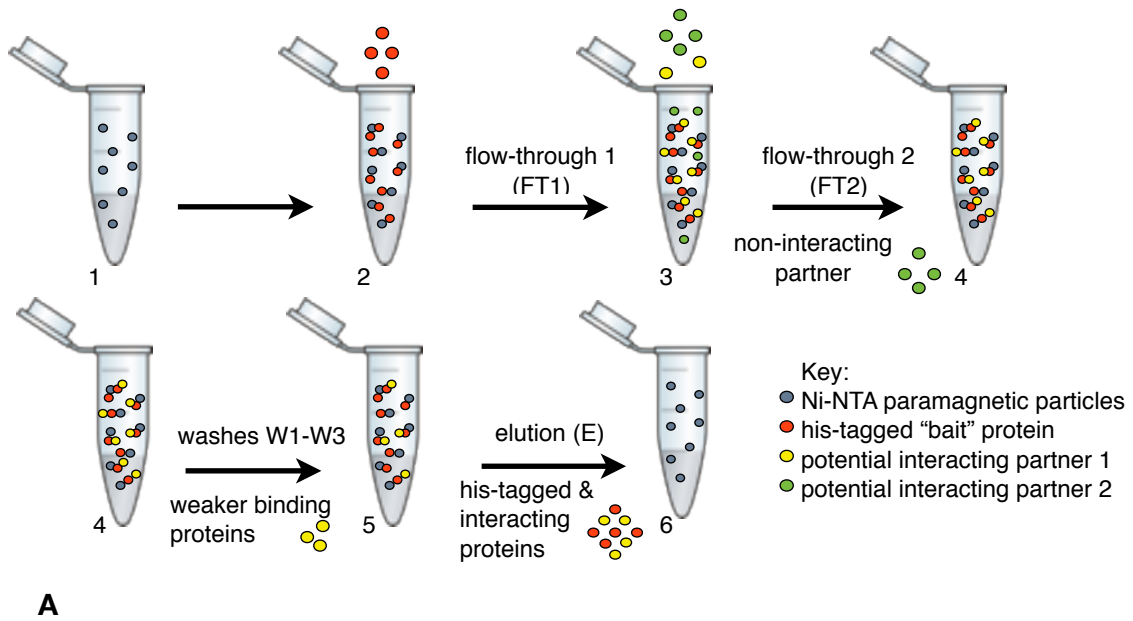


Figure 4.8: Experimental setup for *in vitro* protein interaction assay

(A) The assay procedure and order of addition of protein components is indicated with black arrows and numbering. Collected fractions and the eluted component in each step is illustrated above /below the arrows.

(B) Purified recombinant proteins participating in the interaction experiments.

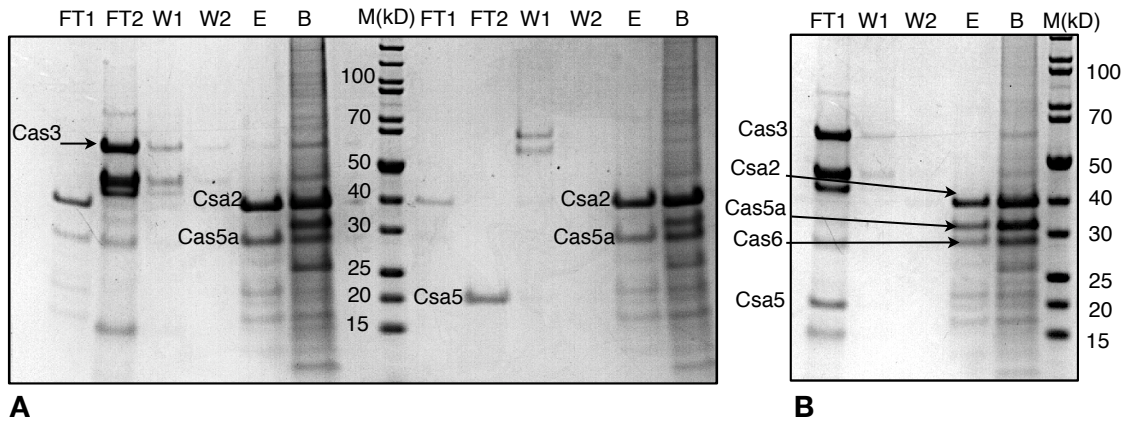


Figure 4.9: Protein interactions of the recombinant Csa2-Cas5a complex

Interaction assay between his-tagged Csa2-Cas5a and Cas3 (first half of gel A), his-tagged Csa2-Cas5a and Csa5 (second half of gel A) and his-tagged Csa2-Cas5a with Cas3, hisCas6 and Csa5 in combination (B). Collected fractions are labelled as (FT1), flow-through after the his-tagged protein was added to the beads; (FT2) flow-through after the interacting partner was added; (W1-2) washes with increasing imidazole concentration; (E) elution with 500mM imidazole; (B) Ni-NTA beads solution after elution heated at 90oC to denature and elute all bound proteins; (M) protein size marker. The smaller molecular weight bands in lanes with Cas3 are degradation products. In gel (B), the presence of both the Csa2-Cas5a complex and Cas6 in the elution fraction is due to the 6xhis-tag on both Csa2 and Cas6 and not a real interaction. Tagged Cas6 was used to investigate whether its presence (\pm RNA) would mediate a bridging interaction between the rest of the assay components, as we were unable to cleave the his-tag off.

4.7 Investigating the properties of the leader sequence of CRISPR locus A

Leader sequences of the CRISPR loci are always present upstream of active CRISPR loci but poorly conserved, and apart from their predicted role in the adaptation stage of CRISPR functioning they also direct the transcription of the locus *in vivo*. In order to determine whether the leader sequence directly upstream of the CRISPR locus A in *S. solfataricus* P2 can act as a canonical promoter we performed *in vitro* transcription studies with CRISPR locus constructs and the *S. solfataricus* RNA polymerase. As leader sequence we considered the genomic region 12333221 (end of gene sso1389) - 12333466 (start of the first CRISPR repeat in locus A). To investigate which elements of the leader sequence are required for transcription two CRISPR locus constructs were generated as described in Materials and Methods, consisting either of the whole leader sequence (245bp upstream from the start of the CRISPR locus) or part of it (165bp upstream) and four repeat-spacer units (CRISPR I, II, figure 4.10). The constructs were cloned into pCR2.1 TOPO vector, linearised following digestion with restriction enzyme *HindIII* and used as transcription templates for the SsoRNA polymerase complex as described in Materials and Methods. Reverse complementary oligonucleotide primers for spacers 1 and 2 were used in the primer extension reaction to verify the presence of the desired transcripts.

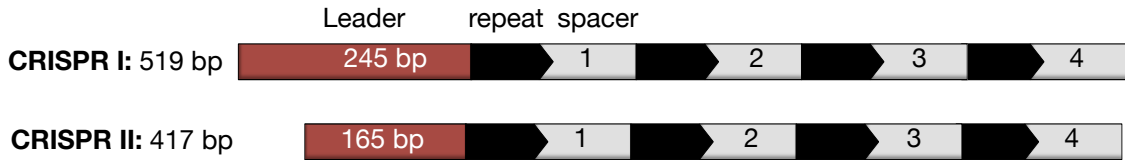


Figure 4.10: CRISPR locus A constructs

Leader sequence is illustrated as a red box, repeats as black arrows and spacers as light grey boxes, numbered in the order they appear in the locus. Sizes are not to scale.

As can be seen in figure 4.11, the CRISPR constructs I and II were transcribed by the SsoRNA polymerase producing fragments corresponding to the length of the repeat / spacer units. The primer complementary to spacer 2 is used as an internal primer when used for the transcripts produced by constructs CRISPR I and II in the primer extension reaction, verifying that the transcript is indeed the series of CRISPR repeats/spacers. Interestingly, the whole sequence (245bp) does not seem to be necessary for transcription, indicating that the transcription factor and RNA polymerase binding sites (BRE box and TATA box) are located within the last 165 bp adjacent to the start of the repeats.

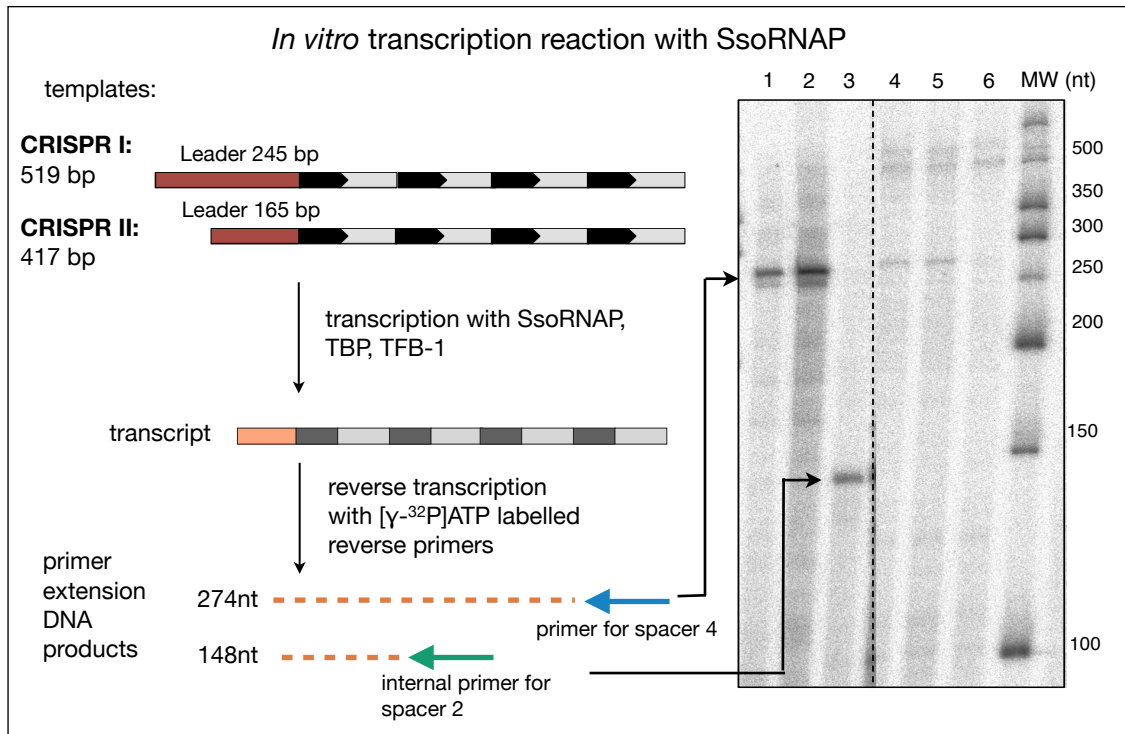


Figure 4.11: *In vitro* transcription of CRISPR locus with SsoRNAP

The left side of the panel contains an outline of the transcription reaction and the synthesised nucleic acid molecules. Colour-coding of leader/repeat/spacer regions as in 4.10. A denaturing 20% PAA gel with the reaction products can be seen on the right side of the panel. Lanes: 1, transcription reaction with template CRISPR I - primer extension with spacer 4; 2, transcription reaction with template CRISPR II - primer extension with spacer 4; 3, transcription reaction with template CRISPR II - primer extension with spacer 2; 4 - 6, negative transcription controls of CRISPR II without TPB, TFB-1 and both TPB/TFB-1 respectively; 6, New England Biolabs low MW DNA ladder.

In order to map the transcription initiation site within the CRISPR leader sequence and gain some insight on the constituent promoter elements the CRISPR constructs used as transcription templates were sequenced by Sanger sequencing. The CycleReader™ DNA Sequencing Kit (Fermentas) was used as described in Materials and Methods. CRISPR II (either as a PCR fragment or as a linearised pCR2.1 TOPO construct) was used as template for the four sequencing reactions with the four different ddNTPs and the reverse complement oligonucleotide for spacer 2 was used as a primer for the kit's thermostable *Taq* DNA polymerase. The reaction products were analysed on a denaturing 15% PAA / 7 M urea gel, and the sequencing information is obtained by comparing the length of the original transcript by the same primer with the specifically terminated strands and "reading" the reverse complement sequence on the primer-extended products. From the sequencing results in figure 4.12 we can deduce that the transcription starts at position -22 of the leader sequence (where 1 is the first nucleotide of the first repeat) and the highlighted sequence (red) before the first repeat is transcribed:

```
5' -GATAAAGAGAAAACCGGTTAAGTTCGTTTTTCATGAAGTTGTTTAAAAGTGTGAAAGTTCGAGTC
TCAATGCGACCGAAACGAATCTTTCTATAATAATTGAACGTTTATAAATGATAGGGTGTATTTCAAT
TTAACATAAAAATCCTTGCGACCAGAAATTGTTAAATTAATTACAACATAAAATTGGTCGCATGAAGAG
TAAAGGGTAGTCATGAAGA TTTATAA GTAAGAAAAGAGAAAGAAAGA
TAGGAAGTATAAAAAACACAACA - 3'
```

Putative BRE and TATA motif sequences are located approximately 21 bp upstream from the transcription start site (highlighted in green and blue respectively), elements which appear to be conserved between leader sequences in *Sulfolobales* (multiple alignment in figure 4.13).

By demonstrating that the CRISPR locus contains a functional promoter sequence and can be transcribed by the *S. solfataricus* transcription machinery *in vitro* we have also established that the transcript we can generate is identical to the native CRISPR transcript produced *in vivo*, and therefore is a valuable substrate to study the primary transcript processing and the mature crRNA biogenesis.

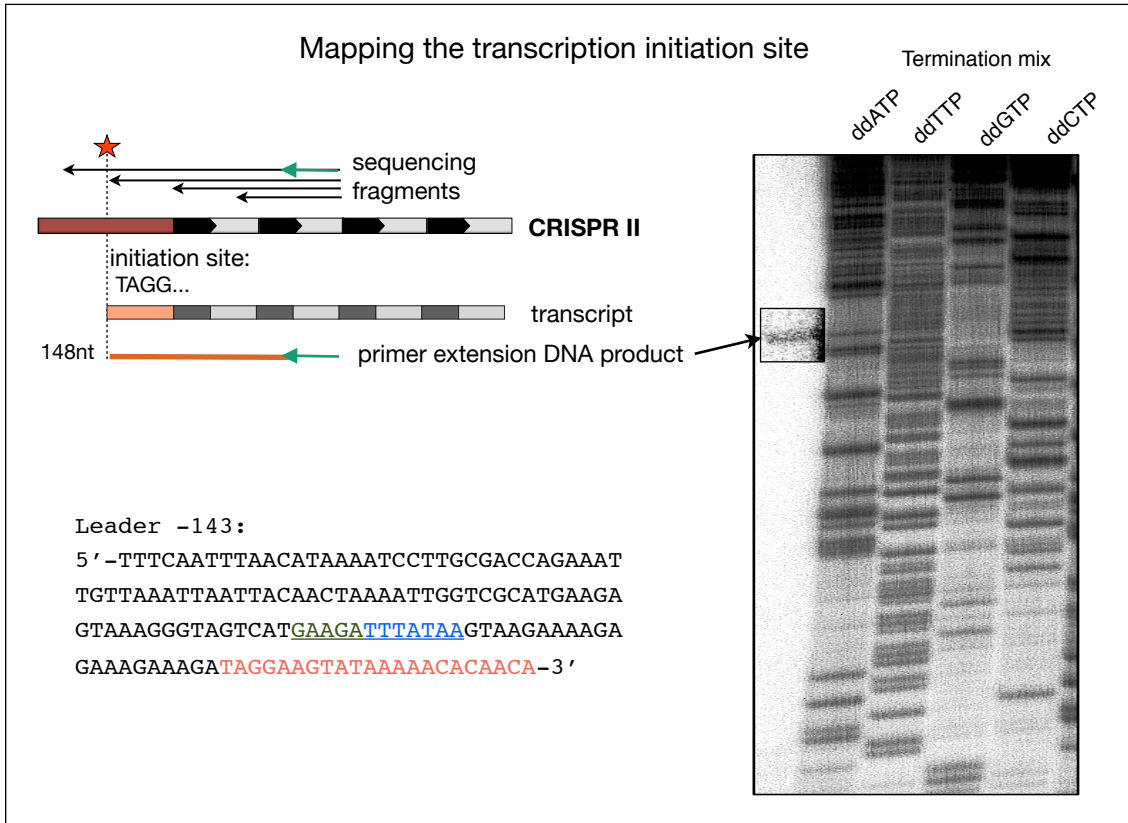


Figure 4.12: Mapping the transcription initiation site

The top left side of the panel illustrates the principle of using Sanger sequencing to map the transcription initiation site. A 15% denaturing PAA gel with the sequencing reaction products is on the right. The first lane of the gel contains the primer extension product from *in vitro* transcription of CRISPR II /spacer2 (see 4.11) and is overexposed compared to the rest of the gel as the product signal was weak. Lanes 2-5 contain the four sequencing reactions with the four different dideoxynucleotide termination mixes.

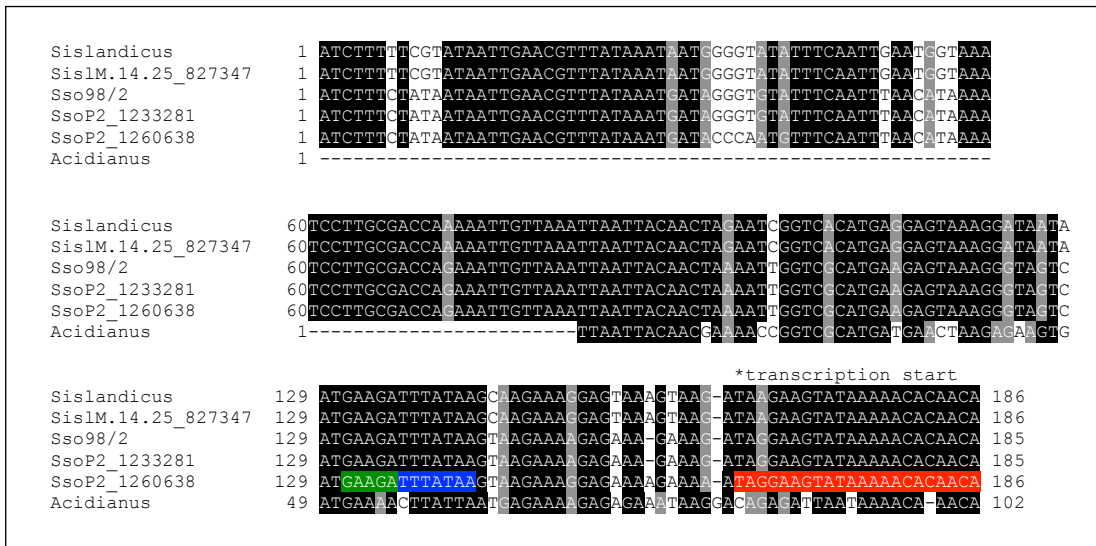


Figure 4.13: Multiple sequence alignment of leader sequences of *S. islandicus*, *S. solfataricus* 98/2, *S. solfataricus* P2 CRISPR A and B and *Acidianus hospitalis*.

Transcription start site in *S. solfataricus* cluster B highlighted in red, putative TATA and BRE boxes in blue and green respectively. Strictly conserved bases highlighted in black. Alignment by ClustalW, shading by BoxShade server.

4.8 Recombinant SsoCas6 is able to process the precursor CRISPR locus transcript and generate mature crRNA fragments

S. solfataricus encodes four Cas6 orthologues (Sso1381, Sso1406, Sso1437 and Sso2004) which belong to the Cas6 I-III superfamily (TIGR01877). Sequence analysis confirms that they contain the signature Cas6 motif within the C-terminal glycine-rich loop characteristic of RAMP proteins, which consists of the consensus sequence GhGxxxxGhG, where h is hydrophobic and the intermediate sequence has at least one lysine or arginine (Makarova *et al.* 2002; Haft *et al.* 2005). The only characterised member of this superfamily, Cas6 from *Pyrococcus furiosus* (Carte *et al.* 2008; Carte *et al.* 2010; Wang *et al.* 2011) belongs to a different family (Cas6 I-A, Makarova *et al.* 2011) and COG (PfuCas6 belongs to COG1853 while all the SsoCas6 proteins to COG5551), and therefore share negligible sequence similarity with SsoCas6 (16%), limited to the C-terminal motif. Structure based threading by the Phyre server detects PfuCas6 as the closest structural neighbour with a 95% sequence coverage suggesting that SsoCas6 also comprises of a duplicate ferredoxin-like fold (figure 4.14), although the catalytic triad identified in PfuCas6 is only partially conserved (Carte *et al.* 2010, Wang *et al.* 2011).

For this reason, the putative nuclease activity of the SsoCas6 orthologues had to be confirmed. The genes *sso1437* and *sso2004* were amplified from *S. solfataricus* P2 genomic DNA, cloned in pDEST14, expressed in *E. coli* and purified by Dr. Shirley Graham using nickel-chelate and gel filtration chromatography. The ribonuclease activity of SsoCas6 was tested against radiolabelled RNA substrates comprising of the CRISPR repeat of locus B and an *in vitro* transcript of the first two repeat-spacer units of CRISPR locus A. The transcript was generated by T7 *in vitro* transcription and nuclease assays were carried out as described in Materials and Methods. 1 μ M of recombinant Sso2004 was incubated with the 152 nt CRISPR transcript for 30 min at 45°C in reaction buffer (20 mM Tris-HCl pH 7.0, 100 mM potassium glutamate, 0.5 mM DTT, 5 mM EDTA) and cleavage products were analysed on a denaturing 20% PAA, 7 M urea gel. SsoCas6 exhibited metal independent ribonuclease activity, and analysis of the cleavage pattern suggested that cleavage occurs at a single site within the repeat sequences of the transcript, yielding the products illustrated in figure 4.15A. Csa2 and Cas5a (individually or in complex) did not exhibit any nuclease activity in the absence of Cas6, nor did they affect the activity of Cas6 or modify the cleavage pattern when Cas6 was present (figure 4.15 A).

To confirm that SsoCas6 cleaves the CRISPR repeat sequence at a single position and determine the cleavage site, 1 μ M of recombinant Sso2004 was incubated with 100 nM of a 25 nt oligonucleotide ssRNA CRISPR repeat sequence with a 15-U 5' extension for 30 min at 45°C, and cleavage products were analysed on a denaturing 20% PAA, 7 M urea gel (figure 4.15 B). By running an alkaline hydrolysis

ladder of the substrate alongside the reaction, the cleavage site was mapped 8 nt from the 3' end of the repeat sequence, at the position indicated by the asterisk:

5' - (15U) - GAUUAAUCCCAAAGGA*AUUGAAAG - 3'

The crRNAs generated by this process are composed of the 5' 8 nt psitag (or 5' handle), the virus-derived spacer sequence, and the 17 nucleotides remaining of the repeat at the 3' end (3' handle) (figure 4.6 E), identical to the cleavage pattern obtained by the PfuCas6. This product is identical to the crRNA fragments extracted from the native aCASCADE, and combined with the fact that SsoCas6 is physically associated with the aCASCADE complex suggests that SsoCas6 is the primary CRISPR transcript processing endonuclease *in vivo*. The Csa2-Cas5a complex did not exhibit any endonuclease activity against the CRISPR transcript in the absence of Cas6 (figure 4.15A), providing further confirmation that SsoCas6 is responsible for the biogenesis of the mature crRNAs found in the effector complex. It is important to mention the analogy with the *E. coli* CASCADE where the processing endonuclease CasE (Cse3) is a subunit of the effector complex.

It is hard to speculate about the reaction mechanism of SsoCas6 and potential catalytic residues without an available structure. In the Phyre-generated structural model, the catalytic triad of PfuCas6 (Tyr31, His46, Lys52 Carte *et al.* 2008) has been replaced by a glutamine at position 32, a threonine at position 45 and a conserved lysine at position 51. Most of the Cas6 conserved residues among the *Sulfolobales*, as identified by multiple sequence alignments, seem to cluster around the central cleft between the two predicted ferredoxin domains in the model (figure 4.14), suggesting a similar path for the RNA substrate as in PfuCas6. A multiple sequence of Cas6 orthologues can be found in Appendix II.

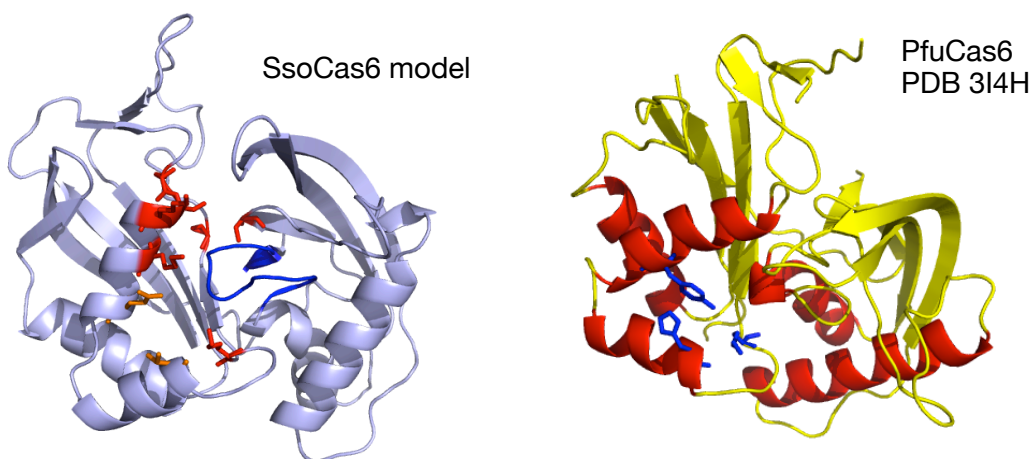


Figure 4.14: Structure of PfuCas6 and model of SsoCas6

The structural model of SsoCas6 was generated by the Phyre fold recognition server using PfuCas6 as template. (*P. furiosus* PDB code: 3I4H) Conserved residues in SsoCas6 are coloured in red, and the signature G-rich loop in blue. Residues corresponding to the PfuCas6 catalytic triad are coloured in orange. The catalytic triad in the PfuCas6 structure is coloured in blue. Images generated with PyMOL.

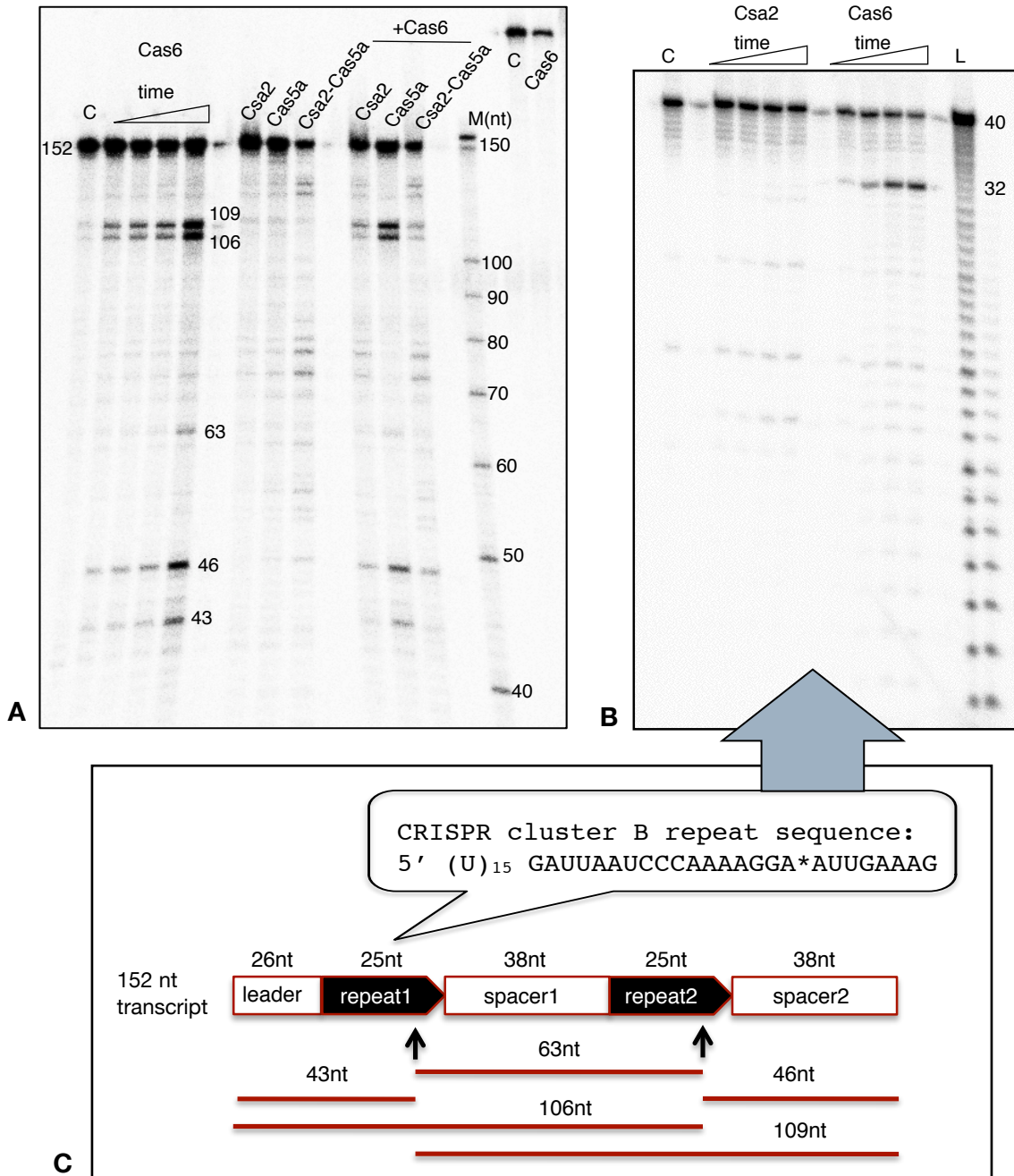


Figure 4.15: CRISPR transcript processing by SsoCas6

(A) A radiolabelled transcript of the first two-repeat spacer units of CRISPR locus B was incubated with Cas6 and other aCASCADE components and products were analysed on a 20% denaturing PAA/Urea gel. Lanes are marked with the respective protein component in each assay. Cas6 is able to cleave the transcript at a single site in each repeat sequence, yielding the pattern illustrated in (C). Csa2 and Cas5a did not exhibit any catalytic activity on the transcript (lanes 6-8), individually or in complex, nor did they alter the cleavage activity of Cas6 (9-11). Cas6 did not cleave a control transcript lacking repeat sequences (lanes 13, 14). C, control reaction without protein; M, RNA Decade marker system (Ambion). (B) Mapping the Cas6 cleavage site within the CRISPR repeat sequence. Substrate is a synthetic RNA oligonucleotide corresponding to a CRISPR repeat with a 15U 5' extension. Csa2 does not cleave the repeat sequence. L, alkaline hydrolysis ladder of the substrate. (C) Schematic illustration of the CRISPR transcript indicating the cleavage sites and the generated products. The cut site within the repeat sequence is indicated with an asterisk. Adapted from Lintner *et al.* (2011).

4.9 The recombinant Csa2-Cas5a complex binds crRNA and forms ternary complexes with target DNA

To determine whether the Csa2-Cas5a complex could bind mature crRNA units and utilise them as guides to recognise and target viral DNA in analogy to the *E. coli* CASCADE, electrophoretic mobility shift assays were carried out to assess its affinity for crRNA and DNA. The substrates used were based on the sequence of the first spacer of *S. solfataricus* CRISPR locus A (A1) and their sequences can be seen in table 4.3.

Name	sequence	notes
crRNA-A1	5' – AUUGAAAGGAACUAGCUUAUAGUUUAGAAGAAAACAAACAA AUAUU GAUUAUCCCAAAA	60nt synthetic crRNA
crRNA-A1_Δ3	5' – AUUGAAAGGAACUAGCUUAUAGUUUAGAAGAAAACAAACAA AUAUU	40nt crRNA minus 3' handle
crRNA-A1_Δ5	5' – GAAAGGAACUAGCUUAUAGUUUAGAAGAAAACAAACAAUA AU GAUUAUCCCAAAAAGGA	60nt crRNA minus part of 5' psitag
A1P control	5' – AGGGUAUUUUUGUUUUUCUUCUAAACUAUAAGCUAGU UC	43nt control RNA
tA1f +PAM	5' – TAATACGACTCACTATAGGGT ATTATTTGTTTGTTCCTTCTAAACTATAAGCTAGTTC <u>TGG</u> AGAGAAGGTG	72nt target DNA +PAM
tA1r +PAM	5' – CACCTTCTCT <u>CCA</u> GAACTAGCTTATAGTTTGTAGAAGAAAACAAACAAATAAT ACCCTATAGTGAGTCGTATTA	72nt reverse target DNA +PAM
tA1f -PAM	5' – TAATACGACTCACTATAGGGT ATTATTTGTTTGTTCCTTCTAAACTATAAGCTAGTTC <u>CCC</u> AGAGAGGTG	72nt target DNA - PAM
RNA_tA1f	5' – AGGGUAUUUUUGUUUUUCUUCUAAACUAUAAGCUAGU UC UGGAGA	49nt target RNA
crRNA native	5' – AUUGAAAGGAACUAGCUUAUAGUUUAGAAGAAAACAAACAA AUAUUAUUAUCCCAAAAAGGA	63nt native mature crRNA

Table 4.3: Synthetic oligonucleotides used in chapter 4

Spacer sequences are highlighted in red, the 5' psi-tag sequence in crRNA is in blue and the PAM in DNA oligonucleotides is underlined.

4.9.1 Substrate analysis of crRNA

Firstly the affinity of the Csa2-Cas5a complex for crRNA was investigated by incubating increasing concentrations of the complex with 100 nM radiolabelled synthetic crRNA oligos at 55°C for 10 min in binding buffer (20 mM MES pH 6, 50 mM potassium glutamate, 0.5 mM DTT, 5 mM EDTA, 5% glycerol) prior to separation by native 10% PAA gels. In order to determine whether the complex binds the crRNA in a sequence specific manner and recognizes elements of the mature crRNA sequence such as the 5' psitag or the 3' handle, the following RNA substrates were used:

- i) crRNA-A1: 60 nt sequence containing the 8 nt 5' psitag, the spacer A1 sequence and 14 nt of the 3' handle. This substrate represents the mature crRNA found *in vivo*, although due to the size limitations of synthetic oligonucleotides, 3 nt are missing from the 3' end.
- ii) crRNA-A1_Δ3: 40 nt sequence missing the 3' handle.
- iii) crRNA-A1_Δ5: 60 nt sequence containing the full spacer sequence and 3' handle, but missing the first 3 nt from the 5' psitag.
- iv) A1P: 43 nt control ssRNA substrate missing the 5' psitag and the 3' handle.

The assays were repeated in triplicate for each substrate and processed using ImageGauge software. Quantification of the bound and unbound RNA fractions enabled the determination of the binding ratio of the RNA-protein complex as the percentage of bound to total RNA for each protein concentration: $(\text{bound RNA}) / [(\text{bound RNA}) + (\text{unbound RNA})]$. An apparent dissociation constant, K_d , for each substrate was estimated by plotting the binding ratios over the protein concentrations. The estimated K_d values and typical assay images are presented in figure 4.16 and table 4.4. These values should be regarded with caution as this type of assay has relatively low sensitivity and is only accurate for proteins that form stable complexes with nucleic acids. The migration of unstable or transient protein-nucleic acid complexes through the gel matrix can result in complex disassembly, leading to overestimation of the apparent K_d values. In the case of Csa2-Cas5a, which is a large multimeric complex, migration of the RNA-protein complexes was extremely slow and a portion of the samples was held up in the wells of the gel, forcing us to adjust the assay conditions in a way that did not reflect the physiologically relevant state of the interaction. For these reasons, the K_d values mentioned here serve only to as a comparative measure for the affinity of the Csa2-Cas5a complex to the different RNA substrates and do not represent the actual K_d values.

From the results we observe that the complex exhibits an overall affinity for ssRNA, as it is able to bind all substrates with comparable K_d values. It is generally considered that the 5' psitag is responsible for crRNA recognition and binding by the CAS effector complexes, as it is the only sequence present in all cloned crRNA sequences from the native aCASCADE, the *E. coli* CASCADE (Brouns *et al.* 2008) and

the *P. furiosus* CMR complex (Hale *et al.* 2009). We were unable to detect a significant difference in the affinity values exhibited by the recombinant Csa2-Cas5a complex for the substrates tested here. For this reason it is not possible to identify which part of the crRNA is recognised and bound by this minimal aCASCADE core, suggesting that potentially other subunits are required. A synthetic RNA oligonucleotide with a complete deletion of the 5' psitag was not generated therefore it is not possible to make assumptions about specific interactions, but it is possible that the protein maintains base-specific interactions with at least the last 5 nt of the 5' psitag and various bases of the 3' handle. The spacer sequence between the repeat-derived sequences of the crRNA is expected to be exposed and available for screening potential invader sequences by Watson-Crick basepairing, therefore the only interaction with the protein should be via the phosphate backbone. Whether this contributes to the relatively high apparent K_d values is unknown.

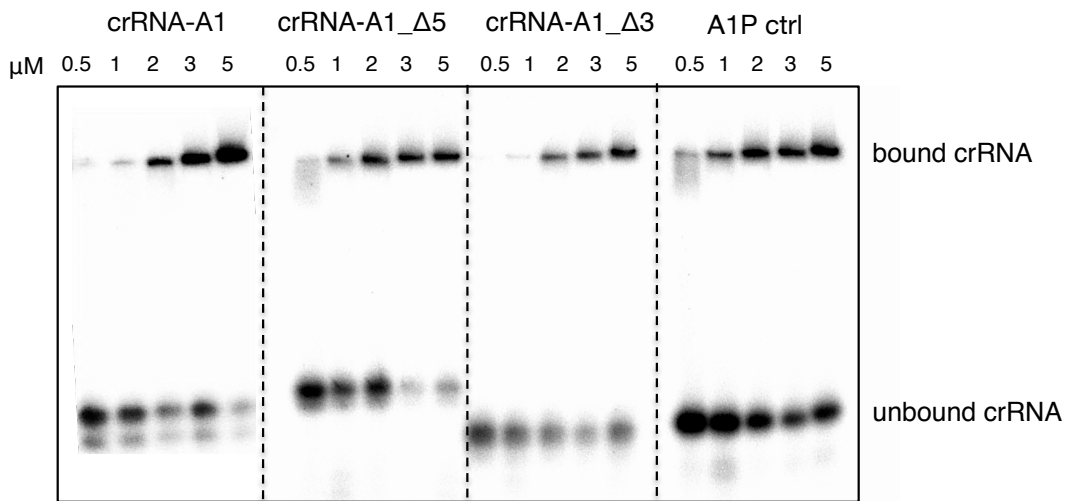


Figure 4.16: Binding of Csa2-Cas5a to crRNA

Areas of photo-stimulated luminescence (PSL), corresponding to uncut substrate and cleaved product, were quantified using Image Gauge software and used to determine the dissociation constants for each substrate as described in the text.

substrate	K _d (μM)	comparative diagram
crRNA-A1	1.5	
crRNA-A1_Δ5	2	
crRNA-A1_Δ3	2.25	
A1P control	2.4	

Table 4.4: Apparent K_d values for Csa2-Cas5a and the various RNA substrates

4.9.2 Csa2 is the main crRNA binding subunit of the Csa2-Cas5a complex

Both subunits of the Csa2-Cas5a complex were tested for their ability to bind crRNA individually, in order to gain insight into the mode of protein - nucleic acid interaction and the role of each subunit in the complex. For this reason, because the Cas5a orthologue found in the native aCASCADE Sso1441 was poorly expressed on its own, the Cas5a orthologue Sso1998 was cloned and expressed in *E. coli* by Dr. Shirley Graham allowing for purification of the recombinant protein with a 6-histidine tag. Electrophoretic mobility shift assays were carried out to compare binding of Csa2 (Sso1442), Cas5a (Sso1998) and the Csa2-Cas5a complex to crRNA-A1 under the conditions described in the previous paragraph. 100 nM of radiolabelled crRNA-A1 were incubated with increasing protein concentrations at 55°C prior to separation on a native 10% PAA gel. Assays were repeated in triplicate, and results are presented in figure 4.17 A & B. Cas5a alone does not bind crRNA, but both Csa2 and the Csa2-Cas5a complex exhibited similar affinities for the crRNA, with apparent dissociation constants between 0.5 - 1 μ M. This observation confirms that Csa2 is the major RNA binding subunit of the complex.

In order to identify potential residues involved in nucleic acid recognition and binding of Csa2, the conserved residue His160 (see 4.2) was mutated to an alanine. The Csa2-H160A mutant exhibited significantly reduced affinity for crRNA compared to the wild type as can be seen in figure 4.17 C, binding only ~10% of the substrate at concentrations where the wild type is capable of 100% binding. We are unable to predict the localisation of the crRNA in the Csa2 structure in the absence of a co-crystal, but the solvent-exposed His160 is obviously involved in the interaction.

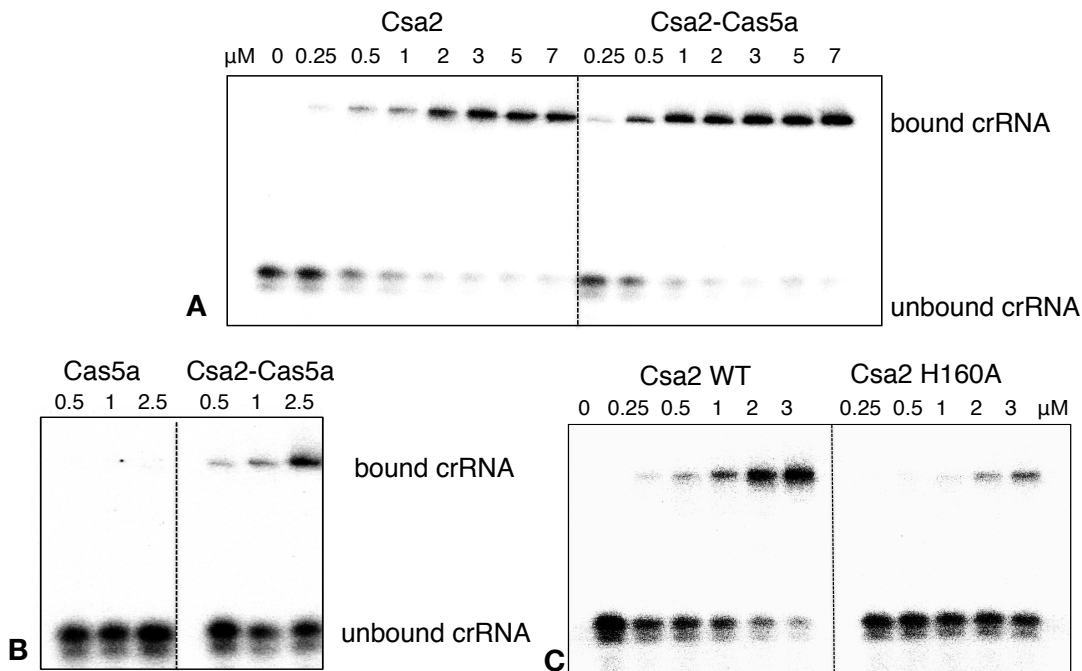


Figure 4.17: Comparative binding of aCASCADE individual subunits and the complex to crRNA

4.9.3 The crRNA - loaded Csa2-Cas5a complex recognises and binds target DNA

The next functional step in the process of viral interference of the *E. coli* CASCADE is the RNA-guided sequence specific targeting of invading DNA, and subsequent recruitment of Cas3 to catalyse the degradation of the invader sequence. The Csa2-Cas5a complex was assayed for similar activity by investigating its affinity for a radiolabelled target DNA in the presence or absence of crRNA. In the absence of crRNA-A1, the affinity of the complex for a ssDNA target was minimal (figure 4.18 A). Nevertheless, when Csa2-Cas5a was incubated with an excess of crRNA, the ribonucleoprotein complex formed was able to recognise and shift a labelled ssDNA complementary to the spacer A1 in the crRNA (substrate tA1f, table 4.3, figure 4.18 A) with an apparent dissociation constant of 750 nM. The recognition is mediated by the basepairing of the central spacer region of the crRNA and tA1f oligonucleotides, resulting in the formation of DNA-RNA heteroduplex and a stable ternary complex with Csa2-Cas5a. The reverse complementary DNA strand (tA1r), containing the spacer A1 sequence, was not gel-shifted by the crRNA-loaded Csa2-Cas5a complex (figure 4.18 A), indicating that the DNA targeting is entirely dependent on the existence of a region of complementarity between the crRNA and the target DNA and the formation of a heteroduplex. Moreover, the crRNA-Csa2-Cas5a complex did not exhibit affinity for a ssRNA target (RNA_tA1f) complementary to the spacer A1 (figure 4.18 B), indicating that the molecular recognition mechanism is specific for targeting DNA, possibly by interactions with the deoxyribose phosphate backbone.

Jore *et al.* (2011) have demonstrated that the molecular mechanism utilised by the *E. coli* CASCADE to recognise invader dsDNA is the formation of an R-loop by basepairing of the protein-bound crRNA with the complementary DNA strand and displacement of the non-complementary strand. The target DNA substrates used in our study were not long enough to observe the formation of an R-loop, although it is possible that a similar mechanism is in operation.

Thus we have demonstrated that the mature crRNAs generated by Cas6 and loaded on the Csa2-Cas5a complex serve as guide RNAs that enable recognition and binding of the invader ssDNA. Predictably we did not observe any cleavage of the bound target DNA by the Csa2-Cas5a complex (figure 4.19). It is hypothesised that in analogy with the *E. coli* CASCADE, accessory CAS proteins are recruited to perform the silencing of the invader DNA. In *E. coli* this role is performed by Cas3, a predicted DEAD-box helicase fused to an HD-nuclease in the *E. coli* CAS system. In *Sulfolobus solfataricus* and other CAS systems these two functional domains comprise different proteins, where Cas3 is a DEAH/X-box helicase always encoded next to a protein with a predicted HD family nuclease domain. It is predicted that both proteins are required to interfere with virus proliferation, but it is unlikely that they interact physically with the Csa2-Cas5a complex as they were not found among the co-purifying proteins

during native expression. Therefore, in order to reconstruct and study the interference pathway for *S. solfataricus in vitro* it is necessary to express these proteins recombinantly.

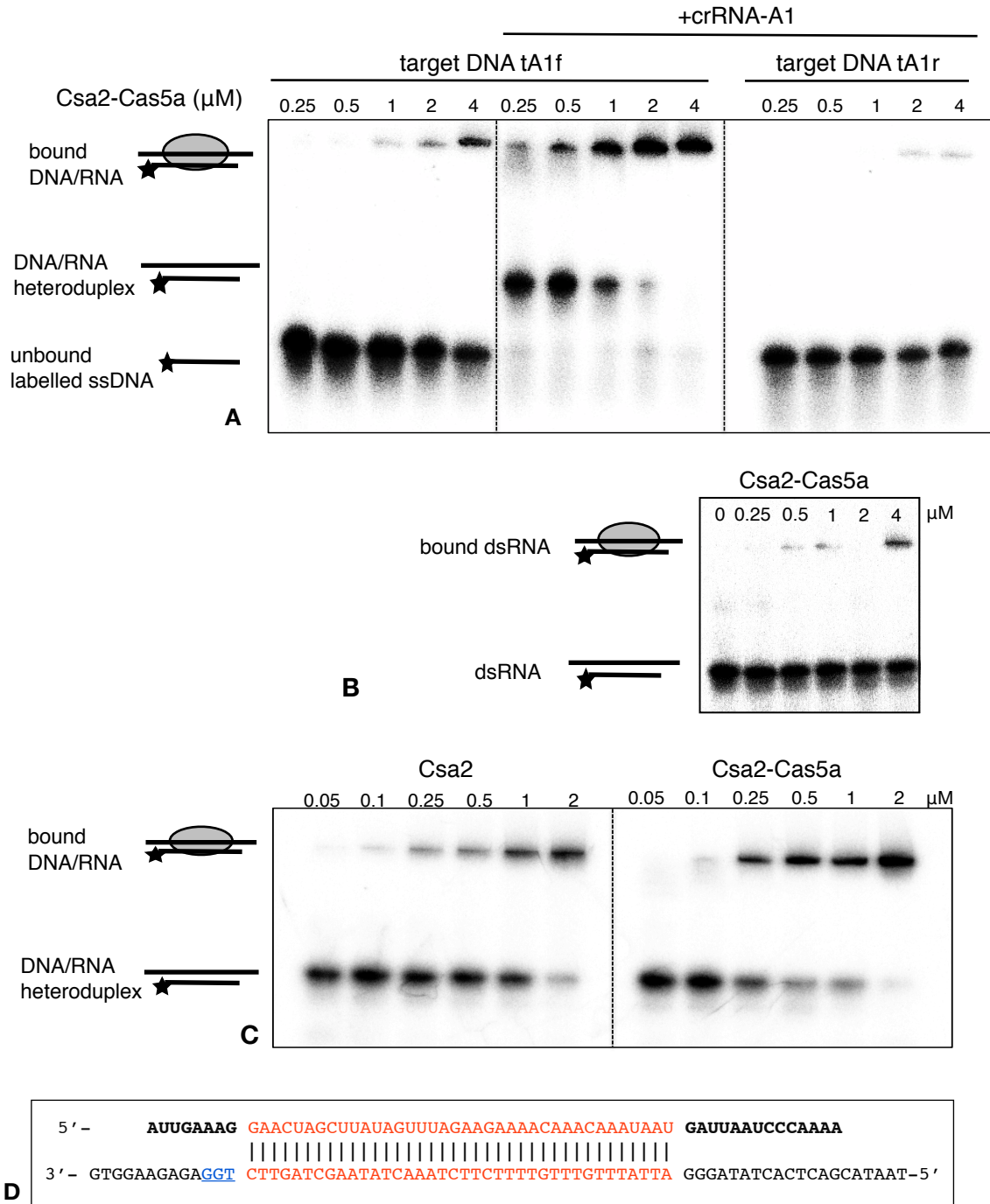


Figure 4.18: cr-RNA mediated binding of Csa2-Cas5a to DNA target

(A) Increasing concentrations of Csa2-Cas5a were pre-incubated with 100nM of unlabelled crRNA for 3 min, and 25nM of labelled target ssDNA were added for 10 min at 55°C. Products were analysed on a native 10% PAA gel. Reactions were repeated in triplicate, and typical assay images are presented here. (B) The Csa2-Cas5a complex showed minimal affinity for an RNA target complementary to the pre-loaded crRNA. Assay conditions and substrate concentrations as for (A). (C) Comparative binding of Csa2 and the Csa2-Cas5a complex to the crRNA/DNA target heteroduplex. Assay conditions as in (A). Both Csa2 and the complex display comparative affinity for the DNA target, with a slightly lower apparent K_d for the complex. This confirms that Csa2 is the main subunit responsible for the nucleic acid

recognition and interactions, with Cas5a potentially involved in alignment and stabilisation of the heteroduplex. (D) The crRNA/target DNA heteroduplex bound by aCASCADE. Spacer sequence in red, the PAM is underlined and highlighted in blue and the crRNA 5' psi-tag is in bold.

As will be discussed in more detail in the following chapter, we were able to obtain a recombinant SsoCas3' but we were not able to express any of the HD nuclease orthologues from *S. solfataricus*, either individually or in co-expression vectors with Cas3. The protein is either highly unstable or extremely toxic for *E. coli*, which leaves native expression in *S. solfataricus* as the only option.

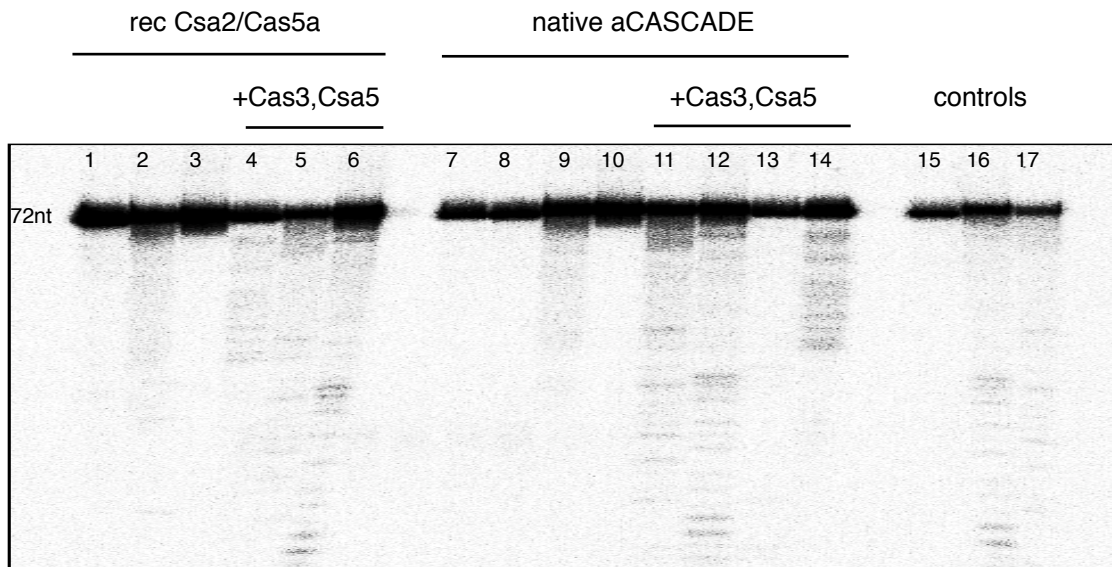


Figure 4.19: Absence of nuclease activity on DNA protospacer targets by aCASCADE

2 μ M of either recombinant or native aCASCADE were pre-incubated with 100nM unlabelled crRNA_A1 for 5 min at 55°C, and the reaction was initiated with the addition of 300nM of ss or ds protospacer DNA targets and 1 μ M of Csa5 and Cas3 where indicated. Reactions were incubated at 60°C for 20min, terminated by Proteinase K treatment for 10min at 37°C and analysed on denaturing 20% PAA/ 7M Urea gel. Protein components are indicated on the gel. In ds substrates, labelled strand is indicated with an asterisk. Lanes: 1, substrate ss tA1f ; 2, substrate ds tA1f/*tA1r ; 3, substrate ds *tA1f/tA1r ; 4-6, as 1-3 ; 7, substrate ss tA1r ; 8, substrate ss tA1f ; 9, substrate ds tA1f/*tA1r ; 10, substrate ds *tA1f/tA1r ; 11, substrate ds tA1f/*tA1r ; 12, substrate ds *tA1f/tA1r, 13, substrate ss tA1r ; 14, substrate ss tA1f ; 15, control ss tA1r ; 16, control ds *tA1f/tA1r ; 17, control ds tA1f/*tA1r. Some specific substrate degradation can be observed with the ss tA1f substrate and the native aCASCADE in the presence of Cas3 and Csa5 (lane 14), and it could potentially be attributed to sub-stoichiometric amounts of the HD nuclease in the partially purified native sample, but this result could not be confidently repeated.

4.9.4 The protospacer adjacent motif (PAM) is not required for target DNA recognition by the recombinant Csa2-Cas5a complex

Sequence analysis of the protospacers corresponding to spacers in CRISPR families of the *Sulfolobales* revealed that they contain conserved dinucleotide motifs at the 5' end, termed "protospacer adjacent motifs" (Lillestol *et al.* 2009). These motifs, as described by Lillestol *et al.*, vary according to the CRISPR family the respective spacer belongs to, leading to the hypothesis that the PAMs may be implicated in the adaptation stage of CRISPR system functioning. Consensus motifs are CC for family I (with tolerance for T at either position), TC for family II and GT for family III. CRISPR loci of *S. solfataricus* P2 belong to families II (clusters A and B) and I (clusters C, D, E, F). *In vivo* experiments where *Sulfolobus* strains were challenged with appropriate protospacer-carrying vector constructs, low transformation efficiencies and high levels of deletions in the transformants were obtained when the predicted PAM motif was present (Gudbergdottir *et al.* 2011). These results indicate that the PAM sequences are required for invader DNA targeting *in vivo*, possibly by mediating a self-non self discrimination mechanism.

The significance of the PAM motif for the functioning of the minimal recombinant system described here was investigated by designing two different ssDNA protospacer targets for crRNA-A1 containing or not the predicted family II PAM motif (CCN). The respective oligonucleotide sequences can be found in table 4.3, termed tA1f+PAM (carrying the sequence TGG) and tA1f-PAM (carrying the sequence CCC) respectively. The affinity of the Csa2-Cas5a complex, pre-loaded with an excess of crRNA-A1, to both targets was comparable (fig 4.20A), indicating that there is no requirement for the presence of the PAM for the recognition of the ssDNA target in the minimal recombinant system assayed here.

To investigate the possibility that accessory aCASCADE proteins are implicated in the PAM recognition, we considered the role of the co-purifying protein Csa5. The gene encoding for Sso1443 (Csa5) was cloned and expressed in *E. coli* by project student Maryam Qurashi, allowing for purification of the recombinant protein through affinity and size-exclusion chromatography. As mentioned before, Csa5 is a small basic 150aa protein of unknown function encoded upstream of Csa2 in a conserved cluster containing Csa5, Csa2 and Cas5a. Its co-purification with the native aCASCADE from *S. solfataricus* indicates a potential functional and physical association. Electrophoretic mobility shift assays carried out by Dr Shirley Graham indicated that Csa5 alone does not exhibit any affinity for RNA or DNA, therefore we investigated whether its presence could alter the affinity of the crRNA-Csa2-Cas5a complex to target ssDNA perhaps by inducing an allosteric conformational change to the proteins. Electrophoretic mobility shift assays carried out in the presence or absence of Csa5 showed no difference in the binding affinity of the crRNA-Csa2-

Cas5a complex to the target ssDNA±PAM, confirming that there is no discrimination at this level (figure 4.20 B). Substrate degradation products observed in high protein concentrations are the result of contaminants in the Csa5 preparation.

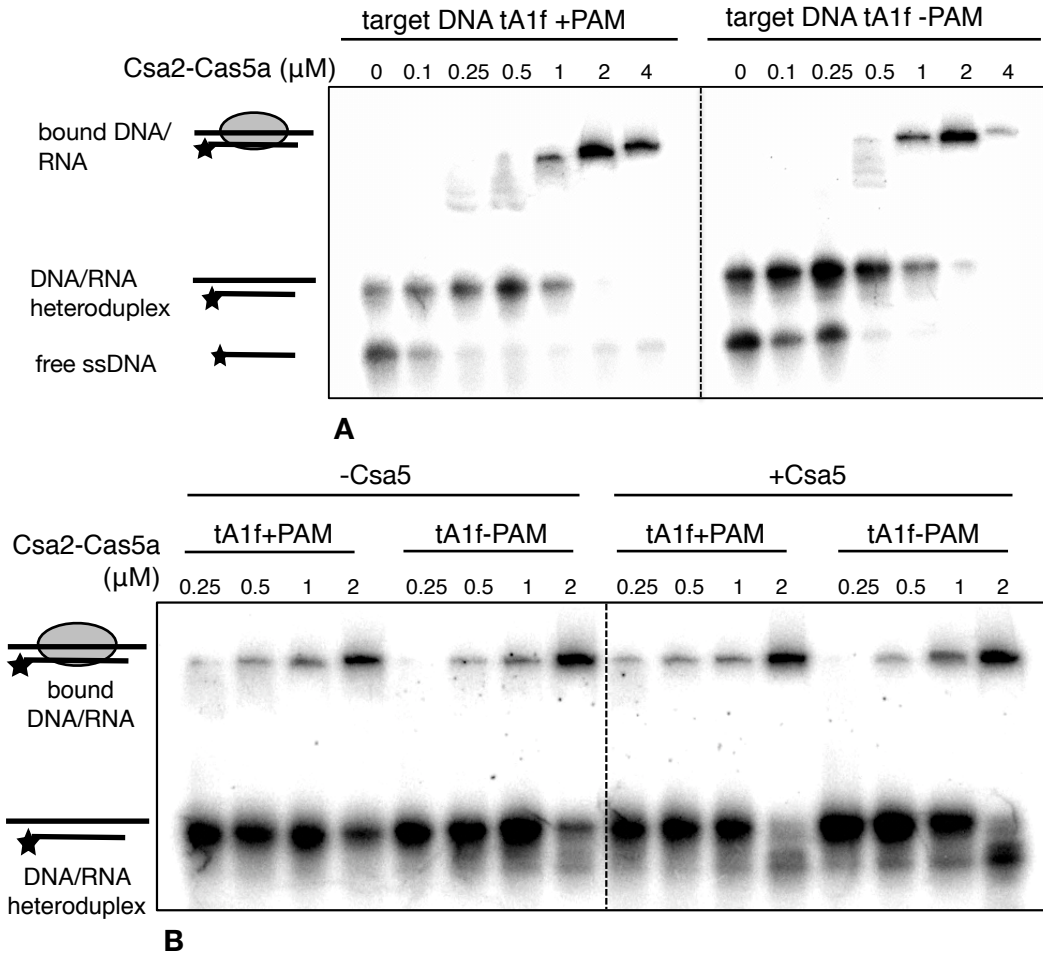


Figure 4.20: Effect of the Protospacer Adjacent Motif on crRNA -guided binding of DNA targets by the aCASCADE.

(A) The Csa2-Cas5a complex displays comparative affinity to complementary ssDNA targets in the presence of crRNA regardless of the existence of a PAM. (B) The Csa2-Cas5a complex was incubated with ssDNA targets ±PAM in the presence of crRNA and in the presence or absence of equimolar amounts of Csa5 in standard assay conditions (55°C, 10min). No effect on DNA binding was observed. The apparent weaker binding of Csa2-Cas5a to DNA in (B) compared to (A) is attributed to the general inconsistent behaviour of this protein in binding assays, depending on the purification batch, the length of time the protein was stored at 4°C and assay conditions. β

4.10 Structural studies and discussion

4.10.1 The structure of Csa2

In order to gain a better understanding of the molecular basis of the archaeal CASCADE the structure of recombinant Csa2 (Sso1442) from *S. solfataricus* was solved by X-ray crystallography by Nathanael Lintner in the group of Martin Lawrence in Montana State University. For full crystallisation conditions, data collection and refinement details, refer to Lintner *et al.* (2011). Phases were determined by multi-wavelength anomalous diffraction on a KAu(CN)₂-soaked crystal which diffracted to 2.0 Å resolution, and by collecting also a 2.0 Å resolution single wavelength dataset from a native crystal. The protein crystallised in space group P2₁2₁2₁ with four copies of Sso1442 in the asymmetric unit. About 8% of the residues in each chain could not be modelled as the electron density was not defined. Structure coordinates were deposited in the protein data bank with accession code PDB ID: 3PS0. The final model consists of four Csa2 chains, but the crystal packing is not thought to represent a biologically relevant quaternary structure as it exhibits closed symmetry, incompatible with the results obtained with TEM (discussed below). The Csa2 monomer consists of three domains arranged vertically, 65 Å in length (figure 4.21 A). The central domain contains essentially an RNA-recognition motif, a ferredoxin-like fold comprised by the four strands of a central antiparallel β-sheet (β6, β7, β1 and β8) and helices α1 and α8, in the βαββαβ topology characteristic of this fold. This RRM motif is extended with a connecting 13-aa loop leading to helix α9, an additional fifth strand to the central β-sheet (β9) and helix α10 which comprises the C-terminus of the protein. Helices α9 and α10 are located on either sides (above and underneath respectively) of the central β-sheet, partially covering the β-sheet surface which is responsible for RNA binding in the typical RRM fold. Moreover, the characteristic sequence motifs containing the aromatic residues responsible for RNA binding are not conserved in the Csa2 central β-sheet (except for Tyr141), which suggests a distinct mode of RNA recognition.

The second and third domains of the Csa2 structure are located in opposite sides of the central RRM-like domain, and are termed “1-3” and “2-4” domains respectively. The former consists of residues 27-46 and 145-180, which form insertions one and three into the RRM domain. Specifically, helix α1 followed by a disordered loop and strands β2-β3 forming an antiparallel hairpin are inserted between the N-terminal β1 and α2, and helix α7 followed by a disordered loop (absent from the model) is inserted between β6 and β7. The “2-4” domain consists of insertions two and four “below” the central RRM domain. In particular, residues 68-136 extending from helix α2 are arranged into short helices α3, α4, α5, α6 and a protruding hairpin composed of antiparallel strands β4, β5, which forms the edge of the crescent-shaped

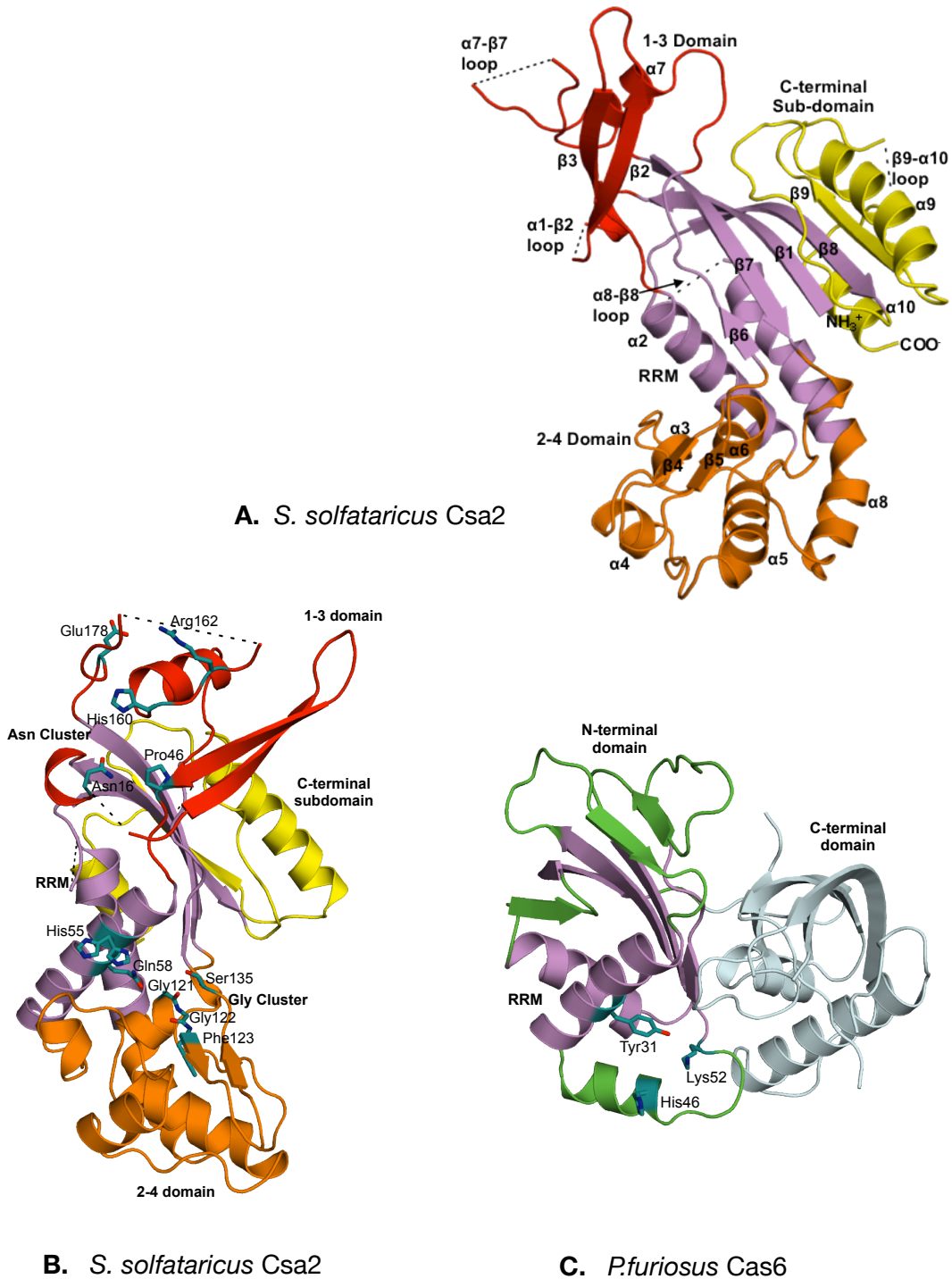
protein. Residues 192-216 form the fourth insertion, which consists of an extended connecting loop between β 7 and the N-terminal half of helix α 8.

Interestingly, the N-terminal ferredoxin fold of *Pyrococcus furiosus* Cas6 is identified by the DALI structural comparison server as the closest structural neighbour of Csa2 (figure 4.21 B & C). The similarity is limited to the RRM domain, with an RMSD of 2.9 Å on 87 aligned residues. As described elsewhere, Cas6 is a metal-independent RAMP superfamily endoribonuclease with a tandem ferredoxin-like fold, with a conserved catalytic triad positioned at the opposite side of the central cleft formed by the β -sheets of the two domains (Carte *et al.* 2008). This similarity contains limited information regarding the function of Csa2, as the respective domains are surrounded by different folds in each protein and the functional sequence motifs are not conserved.

A closer inspection of the conserved residues within the Cas7 superfamily and the Csa2 family (Cas7 type I-A, TIGR02583) in particular, reveals that the majority can be mapped on two clusters on the protein surface, the first on the 1-3 domain and the second at the interface between the RRM and the 2-4 domain (figure 4.21B). The first cluster consists of residues Asn16, Pro46, His160, Arg162 and Glu178 and is referred to as the asparagine (Asn) cluster and the second as the glycine (Gly) cluster and consists of residues His55, Gln58, Gly121, Gly122, Phe123 and Ser135. These clusters form solvent-exposed, basic patches on the concave surface of the protein crescent, and would be suitable candidates for mediating nucleic acid interactions. The glycine cluster in particular is positioned in the “opposite side” of the β -sheet of the RRM domain, reflecting the relative positioning of the PfuCas6 active site with the central cleft (figure 4.21 C). This supports the hypothesis that the conserved Gly cluster plays a functional role, potentially in nucleic acid recognition and binding. Furthermore, we have demonstrated that mutation of the conserved His160 to alanine indeed abolishes the ability of Csa2 to bind crRNA. Additional conserved residues are located in the disordered α 1- β 2 and α 8- β 8 loops (Gly22, Asn23 and Asg240 respectively). The apparent flexibility of the disordered loops and the β -hairpins in the 1-3 and 2-4 domains along with the location of the conserved Asn and Gly clusters suggests their involvement in the recognition and binding of the crRNA, and the subsequent recognition and correct positioning of the target DNA (ss or ds). The nature of the conserved residues is such that it is unlikely that Csa2 exhibits a nuclease activity like Cas6, which was confirmed experimentally. With the RNA-binding surface of the RRM domain partially covered, it is difficult to suggest a path for the crRNA. It was demonstrated however by RNase protection experiments (Lintner *et al.* 2010) that the crRNA is protected completely, indicating that it is bound by the protein throughout its length.

Figure 4.21: Structure of SsoCsa2

(A) Cartoon representation of the Csa2 structure, illustrating the topology and connectivity of the various secondary structure elements. Domains are coloured as follows: RNA-Recognition Motif (RRM) in violet, C-terminal extension of the RRM motif in yellow, 1-3 domain in red and 2-4 domain in orange. (B) Location of conserved residues of the Csa2 family on the Csa2 structure. (C) Cartoon representation of the structure of Cas6 from *Pyrococcus furiosus*. The RRM domain which exhibits similarity to the Csa2 RRM domain is depicted in violet to enable comparison with the respective domain in (B). The location of the conserved catalytic triad is marked with sticks. Adapted from Lintner *et al.* (2011).



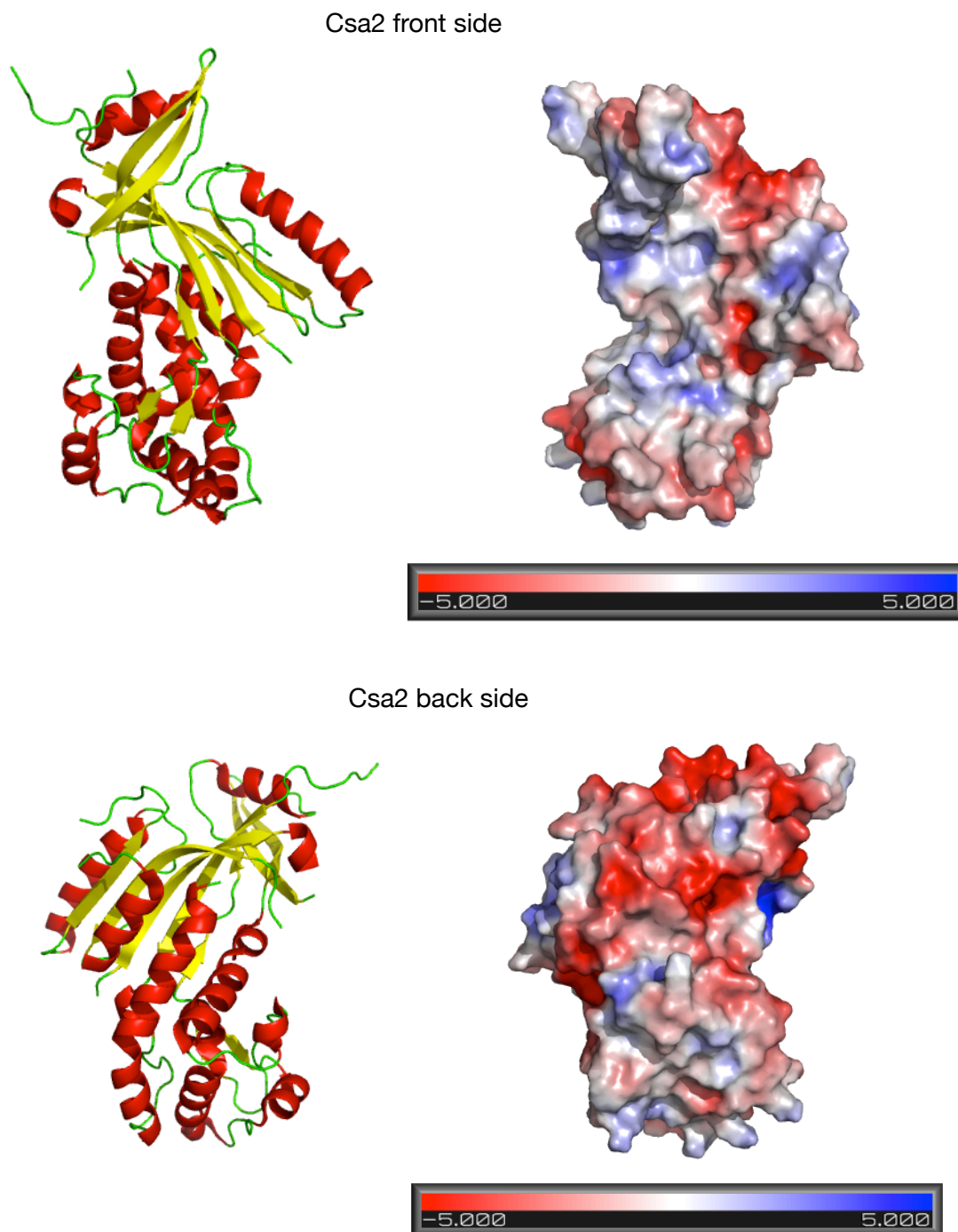


Figure 4.22 : Electrostatic surface map of Csa2

Cartoon representations of the two facets of Csa2 and their respective electrostatic surface map, calculated with APBS tools in PyMOLX11. Surface electrostatic potential is set at ± 5 kT/e and colour gradient is red (acidic) to blue (basic). We can observe a large negative patch on the back side of the protein (non-concave side) corresponding mainly to the 1-3 domain that could mediate nucleic acid interactions.

4.10.2 Arrangement of the native aCASCADE and mechanistic implications

Negative stain transmission electron microscopy (TEM) was employed by Nathanael Lintner to visualise the structural organisation of the native aCASCADE sample, purified from *S. solfataricus*. The complex was arranged in right-handed helices with 14 nm pitch, forming protein filaments with 6 nm width and variable length (figure 4.23 A). For complete experimental details and processed images refer to Lintner *et al.* (2011). These filaments were observed only in the presence of Cas5a and crRNA, while Csa2 alone was shown to be predominantly monomeric/dimeric in solution. The excess of Csa2 over Cas5a in both the native and recombinant purified samples is comparable to the over-representation of CasC in the *E. coli* CASCADE, where it forms a semicircular “backbone” with 6 subunits. It is possible therefore, that the primary component of the helices is Csa2, with Cas5a and the other accessory co-purifying proteins (Csa5, Cas6, Csa4) serving to stabilise or control the nucleation and growth of the complex. The variable length of the helical assembly could also explain the inconsistent behaviour of the aCASCADE complex on the analytical size exclusion column and our inability to estimate the molecular weight. The open symmetry displayed in these assemblies is in contrast to the closed symmetry observed in the asymmetric unit of the crystal, and because the former were observed in the presence of the natural Csa2 protein partner, they are thought to represent biologically relevant arrangements.

Interestingly, it was observed that the length of the Csa2 monomer (65 Å) is comparable to the width of the helix, allowing multiple copies of the Csa2 structure to be modelled onto the helix. Two structural models were proposed by N. Lintner and M. Lawrence to account for the potential functional role of the helical assemblies (figure 4.23 B). In the first model, the extended filaments bind multiple crRNA units and are used to screen target DNA simultaneously, perhaps by wrapping around it and inducing the formation of R-loops as observed for the *E. coli* CASCADE. In the second model, a shorter arch-shaped assembly composed of limited Csa2 subunits is binding a single crRNA, and sub-stoichiometric amounts of Cas5a and perhaps Csa5/Cas6 form the nucleation/termination ends of this partial helix. This second model is reminiscent of the arrangement of the CasC backbone in the *E. coli* CASCADE, and would constitute a more flexible effector complex to patrol the cell for invading DNA, with the added advantage of adjustable length according to spacer length. In any case, it was observed that the groove of the helical assembly is large enough to accommodate either dsDNA or an RNA/DNA hybrid (see Lintner *et al.* 2010).

Our biochemical data suggest that Csa2 is the primary RNA-binding component of the Csa2-Cas5a complex, exhibiting a higher affinity for crRNA over control sequences. However, Csa2 can bind the control RNA substrate suggesting a general sequence-independent binding ability, perhaps as a result of its basic surface

patches. In an effort to analyse the crRNA elements responsible for this specific interaction, it was observed that both the conserved 5' psitag and the 3' handle contributed to the binding specificity, since the protein exhibited comparable affinity for substrates missing one or the other element. Since there is an obvious need for the bases of the spacer sequence to be solvent-exposed and accessible for basepairing with target DNA, sequence-specific interactions must be restricted to the conserved repeat-derived sequences. The 5' psitag sequence seems to be a general feature of crRNAs in all systems studied so far, highlighting its importance as a universal recognition signal of crRNA by the effector Cas proteins. The Csa2-Cas5a complex must also be able to screen invader dsDNA for appropriate targets and stabilise a crRNA/DNA heteroduplex. A crystal structure of crRNA-bound Csa2 (\pm Cas5a) would help elucidate these mechanistic problems.

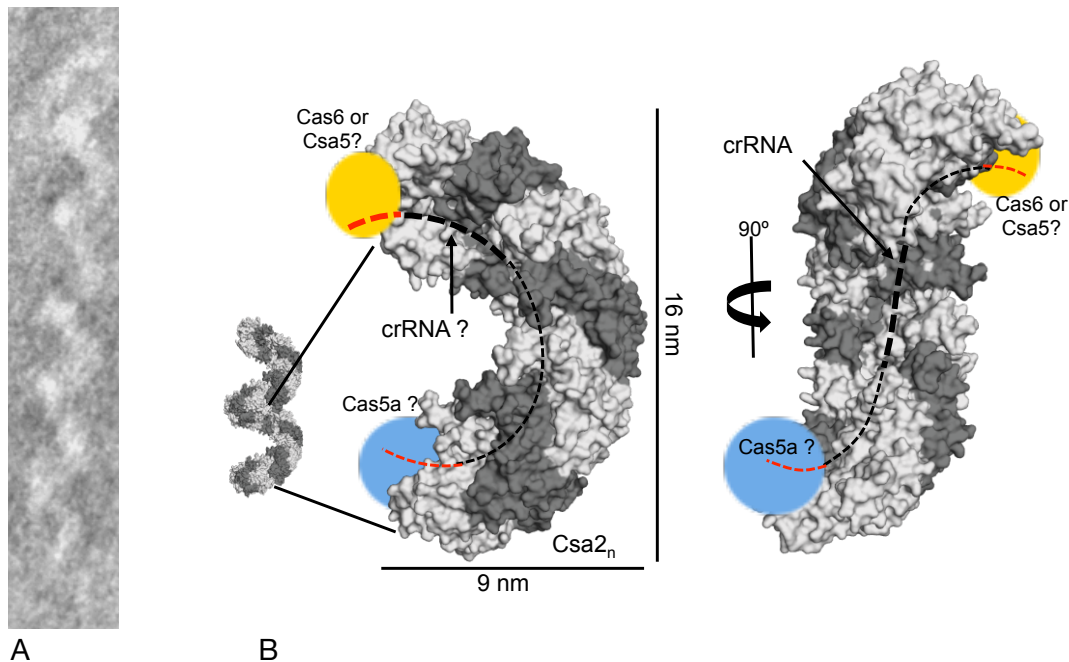


Figure 4.23: Quaternary structural models of native aCASCADE

(A) The right-handed helical structures of aCASCADE visualised by TEM. (B) Model of the potential arrangement of Csa2 monomers in the observed helical assembly, and roles of the accessory subunits. The Csa2 subunits composing the crRNA-supporting core of a partial helix are coloured in alternating dark and light grey. Copies of Cas5a and/or Csa5/Cas6 could be involved in inducing / terminating the polymerisation of Csa2. Adapted from Lintner *et al.* (2011).

In contrast to the observation made by Gudbergsdottir *et al.* (2011) that the protospacer adjacent motif (PAM) is required for *in vivo* targeting in *S. solfataricus*, we do not observe such a requirement in the minimal molecular system identified here. The same conclusion was reached by Manica *et al.* (2011) in their genetic studies of *in*

vivo interference in *S. solfataricus*, where the protospacer of choice did not contain a PAM sequence in its original context. Since the recombinant Csa2-Cas5a complex is lacking some of its *in vivo* partners (Csa5, Csa4, Cas3, HD nuclease) we are not in a position to make definite assumptions about the importance of the PAM motif, especially since it is identified in sequence analyses of the *S. solfataricus* protospacers (Lillestol *et al.* 2009). It cannot be ruled out that one of these accessory proteins (Csa5, Csa4) are responsible for this specific interaction, and perhaps it is required in order for the final step to occur, namely the recruitment of Cas3 and HD nuclease and the final degradation of the target DNA. It could also be the case that the PAM is recognised in a dsDNA substrate, but the molecular basis of such a possibility would require interactions between Csa2 and both the DNA strands simultaneously. It is also possible that this motif plays a key role during the adaptation stage and is essential for the recognition and selection of new spacers, in which case it may be recognised and bound by other Cas components.

4.10.3 Emerging model for CRISPR-mediated interference in Archaea

This chapter describes the first identification and biochemical/structural characterisation of a CASCADE orthologue in Archaea. Comprised by subunits Csa2 and Cas5a (or Cas7 and Cas5 respectively), orthologues of CasC and CasD, the native complex purified from *S. solfataricus* is shown to co-purify with processed CRISPR-derived RNA. A number of transiently interacting proteins also co-purify with the Csa2-Cas5a complex indicating an accessory role, namely Cas6, Csa5 and Csa4. The conservation of these two superfamilies across CRISPR subtypes is indicative of their key role for the structure and function of CASCADE-like complexes, a hypothesis supported by the results presented here. The archaeal CASCADE demonstrates a crRNA - dependent DNA binding activity analogous to the *E. coli* CASCADE, and enables the formulation of a biochemical model for CRISPR - mediated antiviral defence in *S. solfataricus*, which is relevant to all the type I CRISPR subtypes harbouring CASCADE orthologues and type III-B systems with Cmr orthologues (figure 4.24). This model does not describe the adaptation stage in which new spacers are acquired or synthesised, as experimental information for this stage is still very limited and is poorly understood.

In the processing stage of CRISPR functioning, CRISPR loci are transcribed normally by the *S. solfataricus* transcription machinery. In this context, we have demonstrated that the leader sequence directly upstream of the CRISPR locus in *S. solfataricus* acts as a canonical promoter and can direct transcription of the CRISPR locus *in vitro*. Whether some form of transcriptional regulation is taking place is a matter of ongoing research. In the bacterial *E. coli* system, it has been shown that transcription of the both the CRISPR locus and the Cas operon is repressed by H-NS

and derepressed by activator LeuO (Pul *et al.* 2010; Westra *et al.* 2010). In *Thermus thermophilus* the cAMP receptor protein seem to control transcription (Agari *et al.* 2010), while in *S. solfataricus* a novel Cas transcriptional regulator with a binding site for an allosteric effector molecule has been identified in the form of Csa3 (Lintner *et al.* 2011). In all the archaeal *in vivo* systems studied up to now, CRISPR loci seemed to be continuously transcribed and processed (Tang *et al.* 2002; Hale *et al.* 2009; Lillestol *et al.* 2006, 2009), in accordance with a surveillance role in the cell, although there is no information as to whether their transcription is up-regulated in response to an infection.

Subsequently, processing of the CRISPR transcripts into mature crRNA repeat-spacer units is carried out by Cas6 (or equivalent processing ribonucleases like CasE or Csy4), which recognises and cleaves at a single site within the repeat sequences, generating a mature crRNA composed by a 5' 8 nt psitag with the characteristic conserved sequence GAAA(C/G) (Kunin *et al.* 2007), a complete spacer sequence and a less defined 3' handle with the remaining repeat nucleotides. We have demonstrated that the SsoCas6 is a metal-independent ribonuclease that recognises and cleaves specifically crRNA repeats at the single site indicated by the asterisk:

5' - GAUUAUCCCAAAGGA*AUUGAAAG - 3'

Thus, the function of the SsoCas6 is equivalent to that of the euryarchaeal Cas6 from *Pyrococcus furiosus*, even though the two proteins are highly diverged. Repeat sequences in *S. solfataricus* and also in the other subtypes associated with Cas6 are predicted to be unstructured (Kunin *et al.* 2007), therefore the mode of recognition must be sequence-specific as shown for PfuCas6. Members of the Cas6 superfamily are present in CRISPR/Cas subtypes I-A, I-B, I-D and also III-A and III-B. Distinct families are found in I-E (CasE) and I-F (Csy4). Conveniently representatives of each of these three clades have been characterised, revealing the variety of molecular mechanisms utilised by these proteins to recognize and cleave their target, and their co-evolution with the respective repeat types. The fact that Cas6 does not exhibit a stable interaction with the aCASCADE or CMR complex in both systems in which it has been studied reflects this functional versatility and its ability to collaborate with multiple types of effector molecules. Further structural studies of the SsoCas6 are needed to elucidate the molecular mechanism of crRNA recognition and cleavage.

The next step in type I systems (with the exception of type I-D, where there are no CASCADE orthologues) is the incorporation of the processed crRNA in the aCASCADE, the assembly of which seems to be a crRNA-dependent process since the formation of helical aCASCADE assemblies was not observed in the absence of crRNA. The core aCASCADE complex is comprised by Csa2 and Cas5a (Cas7 and Cas5 respectively), with accessory co-purifying proteins including Cas6, Csa5 and Csa4. The complex stoichiometry is undefined, but an excess of Csa2 over Cas5a is

observed, reflecting the abundance of CasC over the other subunits in the *E. coli* CASCADE. Extended right-handed helical assemblies are formed by the Csa2-Cas5a-crRNA complex *in vitro*, but whether this represents the physiological quaternary structure of the complex is unknown. The aCASCADE can recognize specifically ssDNA complementary to the spacer sequence in the crRNA, and form a stable ternary complex with the RNA/DNA heteroduplex. The *E. coli* CASCADE is also able to recognize dsDNA targets via the formation of an R-loop, whereby it displaces the non-complementary strand and enables the basepairing of the crRNA with the target DNA strand. A possible recognition of target dsDNA should also be investigated for the aCASCADE, but was not carried out in the context of this thesis due to time constraints. Cas3 and the HD nuclease are presumed to be recruited to the aCASCADE-crRNA-DNA complex in order to catalyze the final degradation of target DNA. This step has not been biochemically characterised in any of the studied systems so far, but genetic studies in *E. coli* demonstrated that both CASCADE and Cas3 needed to be expressed to produce a resistant phenotype (Brouns *et al.* 2008). A bioinformatics analysis of the *S. solfataricus* Cas3 and HD nuclease orthologues will be presented in the subsequent chapter. Even the presence of a single spacer is shown to be sufficient to confer complete resistance to the respective extrachromosomal element, indicating that this is a rapid and effective mechanism.

The Csa2-Cas5a complex did not recognise RNA targets, but an alternative route is potentially available in organisms that harbour type III-B Cas gene sets, like *S. solfataricus*, *Pyrococcus furiosus* and approximately 60% of the archaea according to the latest analysis by Makarova *et al.* (2011). The Cmr proteins of type III-B systems have been shown to form stable multimeric complexes in *P. furiosus* (Hale *et al.* 2009) and *S. solfataricus* (see chapter 3), which are able to perform crRNA-guided silencing of invader RNAs *in vitro* (Hale *et al.* 2009) as described in detail in Chapter 3. Whether this is the physiological activity of type III-B systems *in vivo* remains to be determined. If this is the case, the co-existence in the same genome of two systems that target DNA and RNA invader elements differentially while sharing the same pool of CRISPR spacers provides an obvious fitness advantage and increases the efficiency of the antiviral defence. The differential processing of the crRNA that associates with the two complexes (aCASCADE and Cmr) represents a type of “labeling” of the crRNA for incorporation into one or the other system. It would be interesting to investigate whether the spacer content of the two crRNA-protein complexes is different, or whether there is a bias towards a specific CRISPR locus family. The latter event would support the suggestion made by Lillestol *et al.* (2009), that individual CRISPR families might exhibit a preference towards specific groups of viruses of extra-chromosomal elements.

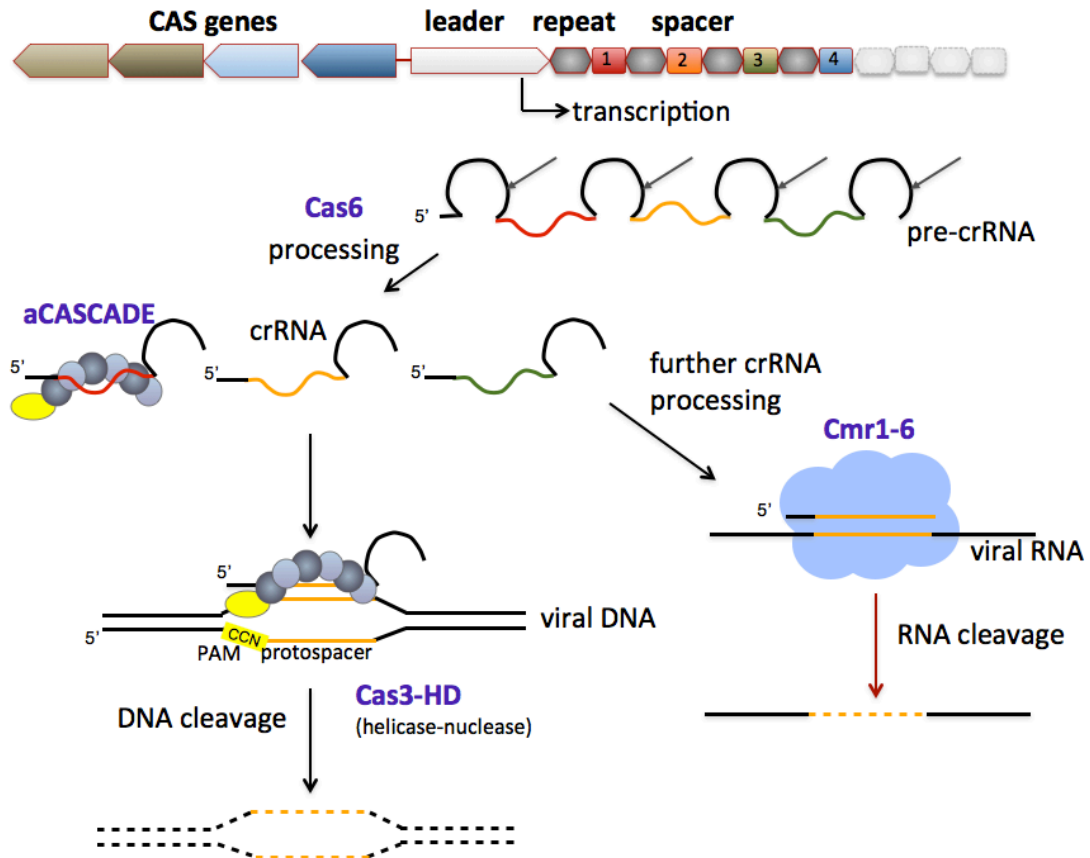


Figure 4.24: Emerging model for CRISPR interference in Archaea

Stages of CRISPR processing and target interference, as deduced by the available experimental data up to date. The pathway on the left involving the aCASCADE would be available to all organisms harbouring type I systems (except perhaps type I-D), enabling the targeting of DNA extrachromosomal elements. The pathway on the right which involves the Cmr complex, would be available to organisms which contain either autonomous type III-B systems or co-existing with other CRISPR subtypes. This pathway would enable the recognition and destruction of any form of invader RNA.

Chapter 5

Initial biochemical characterisation of Cas3' from *S. solfataricus*: a predicted CRISPR-associated helicase

5.1 Introduction

Helicases are a class of enzymes which can effectively couple the free energy derived from NTP hydrolysis to catalyse separation of duplex nucleic acids (NA). They are part of the much larger group of nucleic acid translocases, defined mechanistically by their ability to translocate directionally along nucleic acid strands in an NTP-dependent fashion. This group of enzymes exhibits a high degree of functional diversity and members are known to play key roles in all aspects of cellular nucleic acid metabolism, including genome maintenance, replication and repair, transcription and RNA maturation.

The identification of conserved sequence motifs and several structure-function studies on representative members of these groups have resulted in the classification of helicases-translocases into six superfamilies (Singleton *et al.* 2007). These superfamilies (SF1-6, figure 5.1 focusing on SF1 and SF2) differ primarily in the distribution and primary sequence of up to 11 signature motifs, which are limited to the core domain of the enzymes and provide them with the abilities to: i) bind and hydrolyse NTPs; ii) bind ss- or ds- nucleic acids; iii) convert the chemical energy from NTP hydrolysis to mechanical energy by certain conformational changes (reviewed in Singleton and Wigley, 2002; Tuteja and Tuteja, 2004; Singleton *et al.* 2007; Fairman-Williams *et al.* 2010). Crystallographic studies have revealed that these conserved motifs are divided between two tandem RecA-like domains, at the interface of which the NTP binding pocket is located (reviewed in Caruthers and McKay, 2002; Singleton *et al.* 2007). The strict conservation of these core motor domains indicates that the specific activity of each enzyme family (be it a helicase, translocase or AAA-ATPase) is conferred by non-conserved, modular accessory domains and the potential interactions they mediate. These domains enable activities such as recognition of

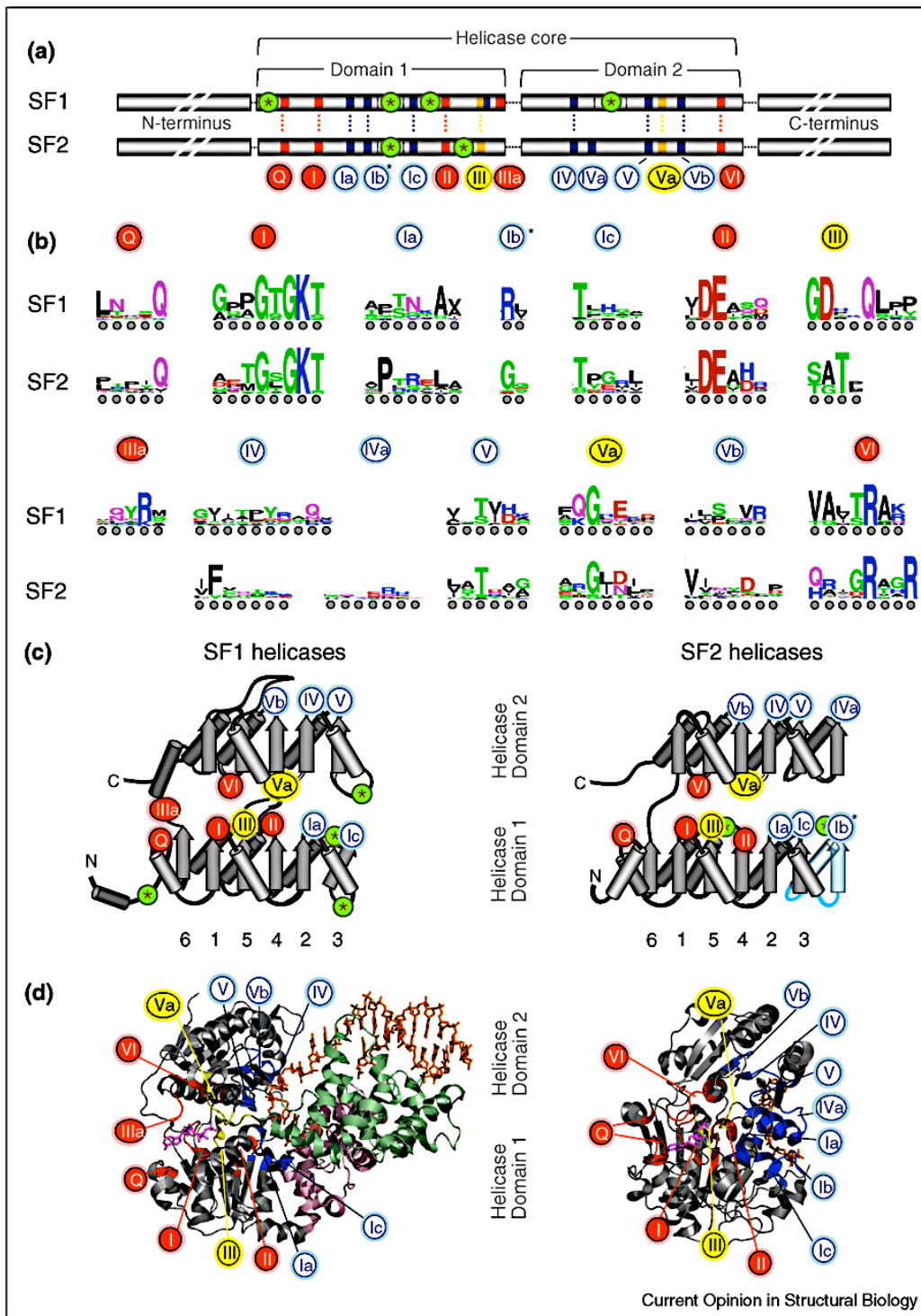


Figure 5.1: Sequence and structural organization of the conserved motifs of SF1 and SF2 NTPases - translocases

(a) Motif and domain organisation of the conserved helicase core focusing on SF1 and SF2. Characteristic motifs are illustrated as boxes, distributed between the two motor domains and colored according to their functional role: red, ATP binding and hydrolysis; blue, substrate nucleic acid binding; yellow, coupling of NTP hydrolysis and NA binding. Green asterisks indicate the typical positions for domain insertions. Distances between motifs are not up to scale. Not all motifs are present in all superfamily members. (b) Consensus motif sequences (c) Topological organisation of the conserved motifs on the secondary structure elements of the RecA motor domains, indicating their spatial proximity and orientation. (d) Motif localisation on representative structures of SF1 (UvrD) and SF2 (Vasa, DEAD-box family). Core motor domains are colored in gray in (b and (c), motifs colored as in (a), accessory domains in light pink and light green. Modified from Fairman-Williams *et al.* (2010)

specific nucleic acid substrates, displacement of proteins or the complementary strand in a nucleic acid duplex and strand annealing. It is beyond the scope of this chapter to provide a full description of the various superfamilies and the specific roles of each motif, but instead we will focus on a phylogenetically distinct group of subfamilies of the SF2 superfamily of helicases-translocases, known as the DExH/D-box protein families of RNA-remodeling proteins.

5.1.2 The DExD/H-box families of RNA-remodeling proteins

These closely related families comprise the majority of the SF2 superfamily proteins, and include almost all known proteins with an RNA-remodeling or helicase activity that take part in various aspects of RNA metabolism (e.g. mRNA splicing, export and degradation, viral replication, miRNA and siRNA processing, transcriptional regulation, translation initiation; reviewed in Silverman *et al.* 2003; Cordin *et al.* 2006; Jankowsky and Fairman, 2007), as well as some DNA helicases/translocases. The individual families DEAD, DEAH, DExH and DExD share between 8-11 of the conserved sequence motifs of NTP-dependent NA translocases/helicases (Jankowsky and Fairman, 2007) and can be distinguished by variations between the motifs or the existence of additional family-specific motifs (e.g. the Q motif, Tanner *et al.* 2003) (figure 5.2). The name of each family derives from the amino-acid sequence of its Walker B motif (motif II), one of the universally conserved signatures responsible for coordinating the Mg²⁺ ion and mediating hydrolysis of the β - γ bond of a bound NTP molecule via the first aspartic and glutamic acid residues (Pause and Sonenberg, 1992; Tuteja and Tuteja, 2004). The other absolutely conserved motifs include motifs I (also known as Walker A) and VI. The Walker A motif (consensus sequence GXXXXGKT/S) is responsible for NTP binding via interaction of the invariable lysine with the β and γ pyrophosphates of the NTP and stabilisation of the transition state during catalysis, while motif VI is exceptionally positioned for energy coupling and interacting specifically with residues in motif II (Cordin *et al.* 2006; Tuteja and Tuteja, 2004; Caruthers and McKay, 2002).

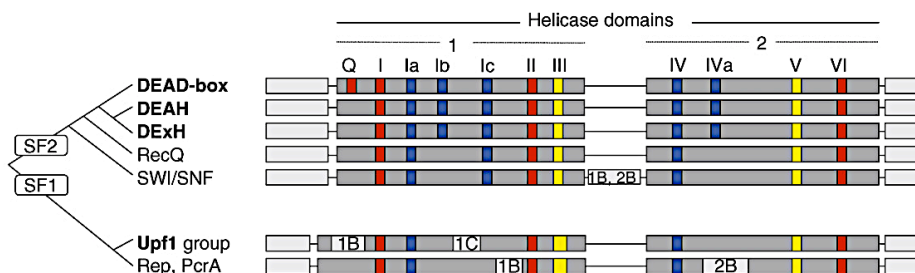


Figure 5.2: Phylogenetic relationships and motif conservation between the SF2 families and between SF1 and SF2 families

Motifs are colored as in figure 5.1. Families that include RNA helicases are in bold. Modified from Jankowsky and Fairman, 2007.

The DExD/H-box group proteins manifest great diversity in terms of size and sequence, with lengths ranging from 400 to more than 1200 residues. They share however a conserved core of ~400 a.a. which contains the canonical sequence motifs. In terms of their structural organisation, crystal structures of DEAD-box proteins have been shown to comprise only of the minimal tandem RecA-like domains and a small N-terminal helix-loop-helix subdomain, with varying interdomain orientations connected by a flexible linker of varying size (e.g. the eukaryotic eIF4a, Caruthers *et al.* 2000; Vasa from *Drosophila*, Sengoku *et al.* 2006). DExH-box proteins in contrast contain auxiliary domains which either physically modulate the activity of the conserved core, mediate protein-protein interactions or are responsible for substrate recognition and specificity. Representatives of this family include the hepatitis C virus NS3 protein and the archaeal Hel308 (reviewed in Pyle, 2008). In DExH-box proteins, the additional domains tend to restrain the movement of the two motor domains keeping the NTP binding pocket in a functional conformation, resulting in the ability to hydrolyse ATP in the absence of nucleic acid. The flexibility of DEAD-box proteins on the other hand, is responsible for the cooperative nature of ATP hydrolysis and NA binding observed in many representatives of this group (Polach and Uhlenbeck, 2002; Yang and Jankowsky, 2005), as the binding of RNA and/or cofactors is necessary to rigidify the interface of the two motor domains and enable NTP hydrolysis (Jankowsky and Fairman, 2007). This crucial difference lies in the root of many observed functional variations between members of the two families (figure 5.3).

In general, DExH-box family members are processive helicases which are able to translocate along ss- or ds- NA (most enzymes tend to be specific for DNA or RNA) and unwind duplexes with a defined 3' to 5' directionality with respect to the loading strand. A single strand overhang to enable loading of the enzyme is usually required. The mode of substrate binding is non-sequence-specific, as the protein interacts mainly with the phosphodiester backbone without distortion of the nucleotide stacking. By contrast, few studied DEAD-box proteins have exhibited unwinding activity, limited to short duplexes (generally below 10 bp) and dependent on the internal stability of the duplex. The family members display varying degrees of sequence specificity, but there is a strict preference for RNA, at least in one of the bound strands (reviewed in Pyle, 2008). The presence of a single stranded NA (typically RNA) usually stimulates unwinding, but not in a single strand-duplex junction as for processive helicases. As demonstrated for the DEAD-box protein Ded1 a single-stranded region in proximity (but not adjacent) to the duplex is used to facilitate the loading of the protein onto the duplex with a yet undefined mechanism (Yang and Jankowsky, 2006). Thus it follows that DEAD-box proteins can also unwind blunt duplexes, although at lower rates. The curious lack of directionality they exhibit when assayed for helicase activity can be explained by their specific mode of substrate binding and unwinding, as elucidated by crystal structures of nucleic acid-bound

DEAD-box proteins like the *Drosophila* Vasa (Sengoku *et al.* 2006). In this structure one of the single RNA strands is severely bent, in a way that prevents canonical basepairing and therefore is implicated in strand separation. This distortion is caused by a helix ($\alpha 7$) on motor domain 1, which adopts a different orientation in processive SF2 helicase structures. It remains unknown whether this is essentially an active mechanism, whereby duplex binding induces the substrate distortion and duplex destabilisation, or the protein simply exhibits affinity for a single-strand transient state of the duplex and prevents reannealing (Sengoku *et al.* 2006; reviewed in Pyle, 2008). A second model for duplex unwinding involves transient conformational changes that take place during ATP hydrolysis and cause duplex destabilisation (Cordin *et al.* 2006; Jankowsky and Fairman, 2007). In any case, DEAD-box proteins appear to have evolved as adaptable “ATP-regulated conformational switches” (Pyle, 2008) which are able to couple energy derived from NTP hydrolysis with a series of catalytic activities depending on the nature of protein partners or specific co-factors.

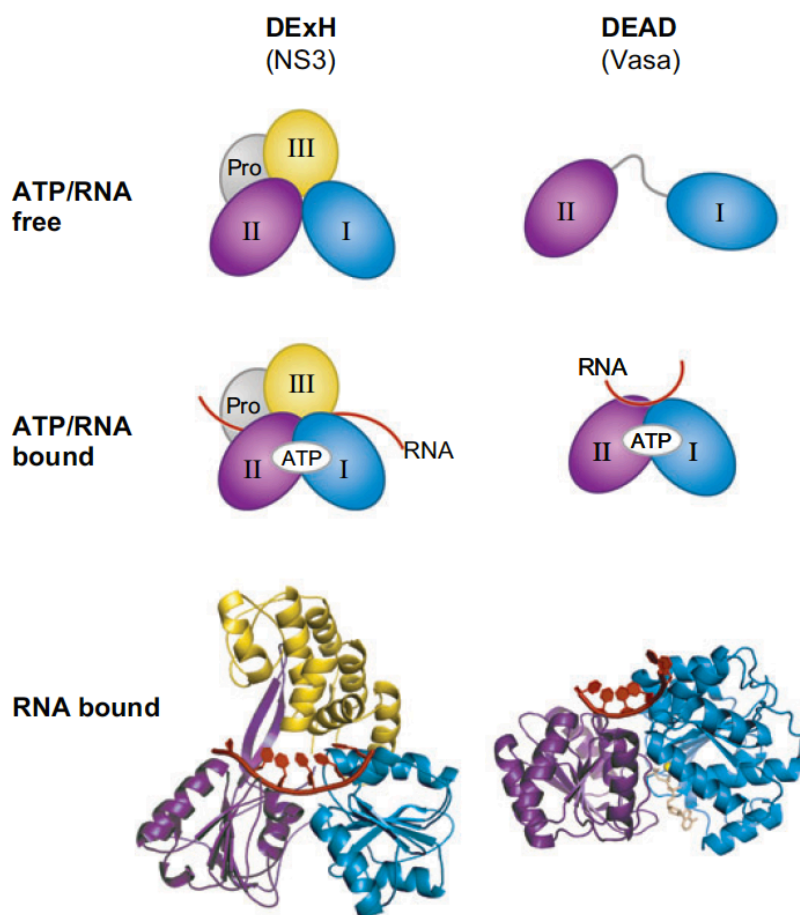


Figure 5.3: Comparison of DExH-box and DEAD-box RNA helicases

The first two rows illustrate the domain organisation in the apo-form and upon ATP/RNA binding of two representative family members, the DExH-box HVC helicase NS3 (left column) and the DEAD-box helicase *Drosophila* Vasa (right column). Ribbon diagrams of the respective crystal structures are presented in the third row. The two motor domains are colored in blue and purple, and the accessory domain of NS3 is in yellow. The bound RNA is in red. Adapted from Pyle, 2008.

Additional functions of the DExH/D-box proteins apart from duplex unwinding include ATP-independent annealing of NA strands (e.g. Ded1, Yang and Jankowsky, 2005; enzymes involved in group I and group II intron splicing, del Campo *et al.* 2009), stabilisation of RNPs or RNA structures (eIF4A; Andersen *et al.* 2006; Pan and Russell, 2011), protein displacement (e.g. Ded1, NPH-II; Jankowsky and Bowers, 2006) or disruption of protein-protein interactions (reviewed in Jankowsky and Fairman, 2007; Pyle, 2008). In all cases, the role of additional domains or interacting proteins is crucial, as DExH-box proteins rarely function individually but instead many of them are integral parts of large multicomponent complexes (e.g. the eIF4F complex or the spliceosome; reviewed in Silverman *et al.* 2003). These protein co-factors act either to stimulate biochemical activity of the DExH/D-box proteins, promote target recognition or regulate them.

Two models have emerged to explain the mechanistic details for ATP-dependent directional ss-translocation of processing DExH-box RNA helicases, alongside the previous “inchworm” and “rolling cycle” models and their variations proposed for DNA helicases (reviewed in Soultanas and Wigley, 2000, 2001). The first is termed the “Brownian motor” model and emerged after observation of the alternating states of NA affinity exhibited by NS3 (the DExH-box hepatitis C virus helicase) at different stages of the ATPase cycle (Levin *et al.* 2005). NS3 is able to bind ss/ds junctions in the absence of ATP, resulting in a directional forward move and duplex disruption. In the second step, binding of ATP would enable a short period of dissociation from the substrate and random “Brownian” movement before the next cycle. The second model, termed the “backbone stepping motor”, assumes a single nucleotide translocation event in every ATP-hydrolysis cycle. The phosphoryl oxygens on the backbone of the tracking strand interact with a conserved threonine in each RecA-like motor domain, the interdomain orientation of which alternates between closed and open states in every ATPase cycle. This results in a step-wise directional translocation along the phosphate backbone of the substrate. Single molecule studies on NS3 revealed that three base pairs are unwound every three cycles of ATP hydrolysis. The mechanism proposed to consolidate this is the following: a conserved tryptophan in the accessory third domain of NS3 is stacked in the ss/ds junction during the 3 nt translocation steps, and its sudden release due to mechanical tension buildup results in duplex unwinding (Myong *et al.* 2007).

5.1.3 The CRISPR-associated putative DExH-box helicase Cas3

A putative helicase was identified among the genes associated with the CRISPR arrays since they were believed to be a novel repair system (Jansen *et al.* 2002; Makarova *et al.* 2002; Haft *et al.* 2005). Alignments of the amino acid sequences of Cas3 proteins revealed the seven conserved signature motifs for proteins of SF2,

with the consensus sequence D-E-X-H in motif II (Walker B) classifying these proteins in the DExH/D family. The helicase core was always fused to or encoded next to a putative HD-nuclease domain, and was initially considered to be the prokaryotic dicer analogue in the context of the novel antiviral system (Makarova *et al.* 2006). Thus, the name Cas3 refers to the whole polypeptide comprising both the predicted helicase and HD-nuclease domains, while when encoded separately they are referred to as Cas3' and Cas3'' respectively.

The HD family of predicted phosphohydrolases was first identified by Aravind and Koonin (1998), who also noticed that they frequently occur as accessory domains of helicases, polymerases and nucleotidyl-transferases, suggesting an implication in various aspects of nucleic acid metabolism. Members of this family are characterised by three strictly conserved (I, II, V) and two less widely conserved motifs (III, IV) including a conserved histidine (I), a histidine-aspartate pair preceded by two hydrophobic residues (hhHD - motif II) from which the family name derives, and a conserved aspartate (V). These motifs are not necessarily close in the primary amino-acid sequence, but they are in close proximity in the tertiary structure of the polypeptide chain, where they are predicted to participate in coordination of a divalent cation necessary for catalysis. The catalytic activity of these enzymes is the hydrolysis of a phosphoester bond in a wide variety of substrates.

According to the most recent classification of the CRISPR/Cas system by Makarova *et al.* (2011), Cas3 is the signature gene for all type I systems. In subtypes I-A and I-B the putative helicase and HD-nuclease domains are encoded separately (*cas3'* and *cas3''*), while subtypes I-C, I-D, I-E and I-F encode one multidomain protein (*cas3*). Studies in the *E. coli* type I-E system provided the first information about its role, when it was discovered that both the *cascade* set of genes and the *cas3* gene were required for the resistance phenotype, but not for the precursor crRNA processing (Brouns *et al.* 2008). Similar observations were made for *cas3* in *Pseudomonas aeruginosa* (Cady and O'Toole, 2011). It was hypothesized that the function of Cas3 (containing both the helicase and HD-nuclease domains in this subtype) would involve degradation of the invader DNA via the HD-nuclease domain and release of the crRNA via the DExH helicase domain (Brouns *et al.* 2008, van der Oost *et al.* 2009; Jore *et al.* 2011).

Even though it was one of the few genes for which a functional prediction could be made, the first biochemical characterisation of a type I-E Cas3 orthologue was published in 2011 by Sinkunas and colleagues. Cas3 from *S. thermophilus* (referred to as SthCas3 in this chapter) was found to possess ATPase activity stimulated by ssDNA regions, ATP-dependent helicase activity and metal-dependent single strand nuclease activity attributed to the HD-domain (Sinkunas *et al.* 2011) Minimal levels of ATPase activity were observed in the absence of NA, but the

presence of ssDNA (and not dsDNA or RNA) greatly stimulated the rate of ATP hydrolysis (maximum rate ~ 38 moles ATP \times moles⁻¹ protein \times min⁻¹). ATPase activity was also supported by GTP instead of ATP and enhanced by divalent cations. In terms of helicase activity, SthCas3 was able to processively unwind dsDNA and RNA-DNA heteroduplexes in an ATP/Mg²⁺-dependent manner, with a 3' to 5' directionality. Conserved aspartate residues in the N-terminal HD-nuclease domain of the SthCas3 were identified as the catalytic residues responsible for the Mg²⁺-dependent unspecific degradation of ssDNA, but not dsDNA. Mutations in the helicase domain did not affect this activity. These results are in agreement with the general characteristics of processive DExH helicases as described by Pyle (2008), except for the apparent lack of ATPase activity in the absence of nucleic acid. The low ATPase rate even in the presence of ssDNA is consistent with the local RNP remodeling roles these proteins play in the various cellular processes they participate in.

Shortly after, Howard *et al.* (2011) reported the purification and characterisation of Cas3 from *E. coli* and *Methanothermobacter thermautotrophicus* (both type I-E systems). An additional activity for Cas3 was discovered, namely that it could anneal DNA and most importantly RNA strands into complementary dsDNA duplexes forming R-loop structures. This activity was dependent on Mg²⁺/Mn²⁺ and an active N-terminal HD-nuclease motif, but independent of ATP. Mutations in the conserved sequence of the latter resulted in reduced R-loop formation, indicating a potential interdomain cooperation. Nuclease activity however was not detected *in vitro*, suggesting that it is not necessarily required for R-loop formation, although the authors do not preclude such an activity *in vivo*. The *E. coli* Cas3 was also able to unwind R-loops in an ATP-dependent manner, and these two conflicting activities seem to be regulated by ATP concentration potentially triggering a conformational change (Howard *et al.* 2011). The levels of ATP hydrolysis were comparable in the presence or absence of ss/ds NA, in line with previous observations about modular DExH proteins (Jankowsky and Fairman, 2007). The observed differences between the biochemical abilities of two Cas3 orthologues of the same subtype (I-E), but from different organisms, illustrate the remarkable versatility of these enzymes and the need to characterise them in conjunction with the large nucleoprotein complex they associate with, namely CASCADE.

The role for Cas3 in the context of CASCADE mediated interference in subtype I-E as proposed by Sinkunas *et al.* (2011), begins with an initial strand separation event on the dsDNA invader resulting in R-loop formation between the crRNA and the complementary protospacer (figure 5.4). Since both the CASCADE-crRNA complex and Cas3 have been shown to promote R-loop formation independently (Jore *et al.* 2011; Howard *et al.* 2011), the exact series of events is unknown. Howard *et al.* suggest that CASCADE could initiate the crRNA invasion into the duplex and then

recruit Cas3 to extend and stabilise it, perhaps by further unwinding of the DNA duplex (Sinkunas *et al.* 2011). Within the R-loop structure, the HD-domain could cleave the displaced unpaired DNA strand at a single site, after which remodeling and unwinding of the crRNA-DNA duplex could take place to enable a second cleavage event in the protospacer strand. The result of this multi-stage procedure would be a double-strand cleavage product within the protospacer sequence of the invader DNA, reminiscent of the pattern observed by Garneau *et al.* (2010) in *S. thermophilus*. Moreover, the observation that CASCADE could recognise and bind to protospacer sequences regardless of the flanking PAM motif let to the proposal of an additional role for Cas3 by Sinkunas *et al.* (2011) in recognising the appropriate PAM motif in target protospacers. In support to this claim is the fact that one of the major functions of accessory domains in DExH helicases is to confer target sequence specificity (reviewed in Pyle, 2008). Given the observed differences in Cas3 biochemical activities between orthologues, it is unknown whether the mechanistic details of this last stage of CRISPR interference would be identical between subtypes, and even between members of the same subtype.

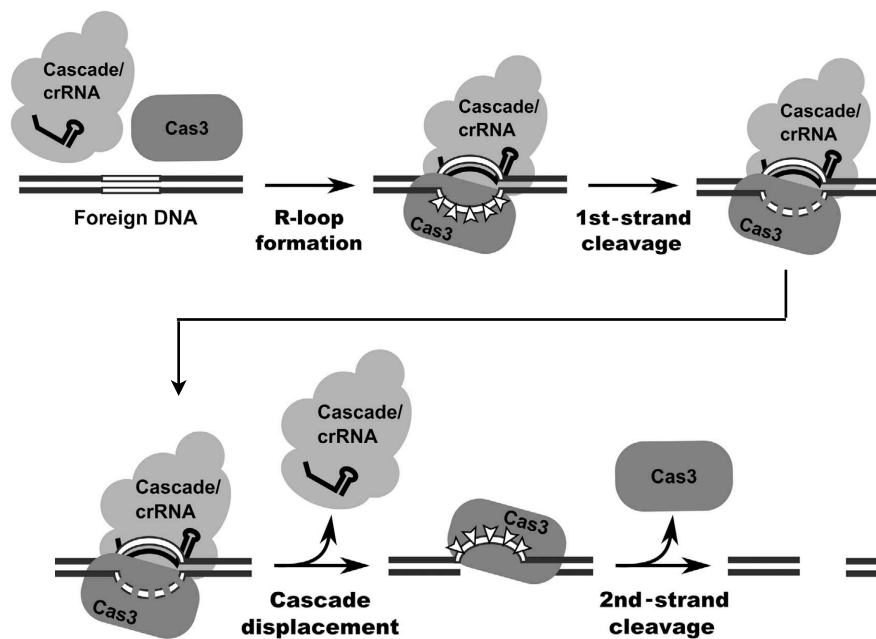


Figure 5.4: Proposed mechanism of action for Cas3 in type I-E systems
Detailed description in text. Adapted from Sinkunas *et al.* 2011.

The separately encoded HD-nuclease (Cas3'') from *S. solfataricus* (gene name: sso2001) was found to demonstrate metal-dependent endonuclease activity on dsDNA and dsRNA substrates, preferentially at G-C base pairs (Han and Krauss, 2009). In contrast, the separately cloned and purified HD-nuclease subdomain of Cas3 from *T. thermophilus* demonstrated nuclease activity against single-stand DNA substrate, in agreement with the activity observed for Cas3 from *S. thermophilus*, the only difference being its stimulation by Mn^{2+} , Co^{2+} , Ni^{2+} and Zn^{2+} instead of Mg^{2+} (Mulepati and Bailey, 2011). The crystal structure of the HD subdomain of TthCas3 was solved by Mulepati and Bailey (2011) and can be seen in figure 5.5. Comparison of this to an available structure for the stand-alone Cas3'' from *Methanocaldococcus jannaschii*, solved by a structural genomics initiative, reveals an overall fold similarity indicating the proteins' phylogenetic relations but also some key differences in topology and arrangement of conserved residues important for catalysis in both proteins (Mulepati and Bailey, 2011). Whether these structural differences are the basis for the distinct substrate preferences exhibited by the two classes of proteins remains to be seen. It should be expected however that since they represent separate evolutionary units Cas3' and Cas3'' might have evolved distinct modes of interaction with each other and with the subtype-specific effector complexes and distinct functional characteristics.

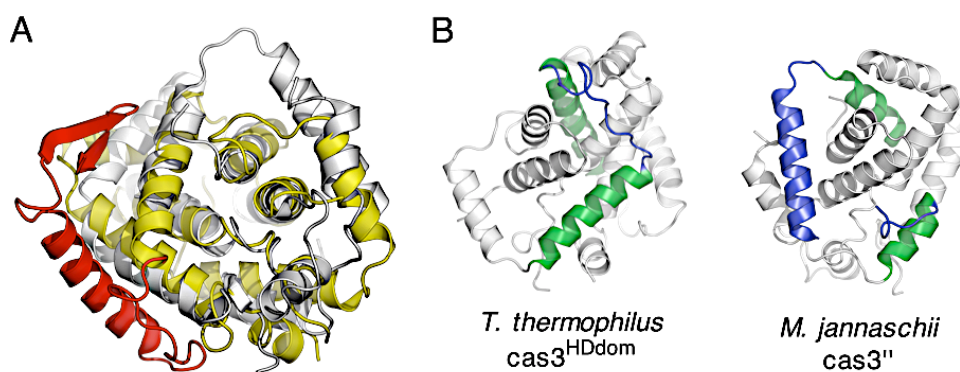


Figure 5.5 Structure of the HD-domain of Cas3 from *T. thermophilus*

(A) Superimposition of the TthCas3 HD-domain (colored in yellow and red) and Cas3'' from *M. jannaschii* (colored white). Structures are largely superimposable, apart from an additional helix-β hairpin in TthCas3 (red). (B) Individual structures of the TthCas3 HD-domain and MjaCas3''. Adapted from Mulepati and Bailey, 2011.

5.2 Cas3' in *Sulfolobus solfataricus*

As mentioned in previous chapters, *S. solfataricus* P2 harbours types I-A and III-B CRISPR/Cas systems and encodes eight putative Cas3' and Cas3'' family orthologues (figure 1.21). This chapter deals with the purification and initial biochemical characterisation of the Cas3' orthologue Sso1440 (referred to as SsoCas3' throughout this chapter). Considering their predicted physical and functional association interdependence, several unsuccessful attempts were made to clone and purify a Cas3'' orthologue in *E. coli*, either individually or by co-expressing it with Cas3'. This could be a consequence either of the protein's structural instability or of the enzyme's toxicity to the host cell. Either way, it is an indication that a protein partner (probably Cas3') is required to provide folding stability and/or regulate its toxic activity.

A close inspection and comparative sequence analysis of the SsoCas3' amino-acid sequence reveals the eight conserved domains characteristic of SF2 helicases. The arrangement and sequence of each motif can be seen in figure 5.6. The sequence of motif II (Walker B) is D-E-F-H, classifying this protein into the DExH-box family.

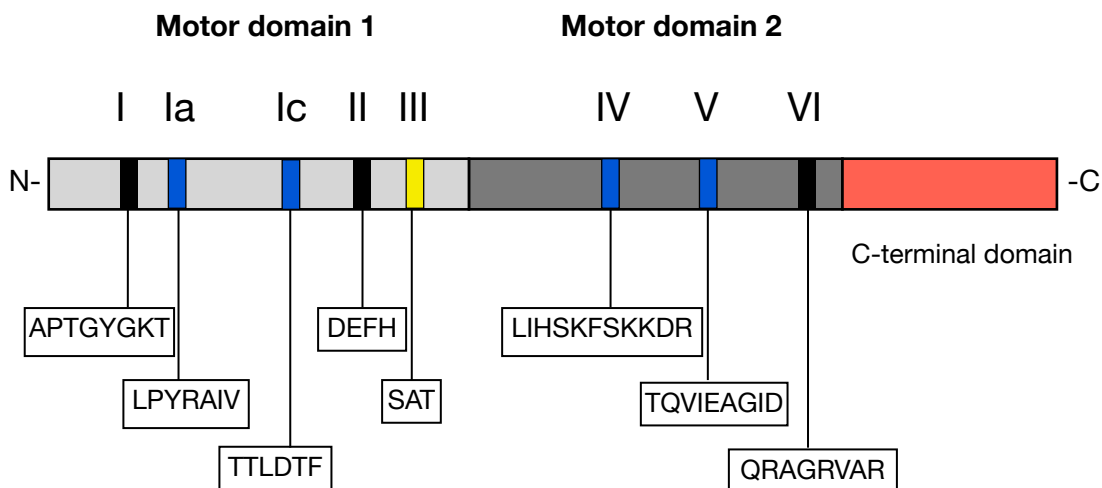


Figure 5.6: Domain arrangement of SsoCas3'

The position and sequence of conserved helicase motifs (coloured bands) is indicated on the putative motor domains of SsoCas3, shown in light and dark grey blocks. Motifs are coloured according to their function: black, involved in ATP hydrolysis and substrate NA binding. The C-terminal domain suggested to mediate protein interactions is shown as an orange block. Distances between motifs are not to scale. Consensus sequence motifs as in Fairman-Williams *et al.* 2010.

The C-terminal 145 residues of SsoCas3' do not contain any recognisable domain although they are conserved within *Sulfolobales*. The protein fold recognition

server PHYRE predicts that this domain is largely helical and models it on the protruding terminal stalk-like domain of another DExH-box helicase, Mtr4, involved in RNA processing. In this protein, the stalk itself does not have a function but acts as a linker for an additional β -barrel globular domain that mediates RNA interactions (Weir *et al.* 2010). It is likely that this C-terminal domain mediates itself protein-protein or protein-NA interactions in SsoCas3'. The model generated by secondary structure threading can be seen in figure 5.7.

Figure 5.7: Structural model of SsoCas3', generated by Phyre2

The two core RecA-like motor domains are colored in yellow, green and blue rainbow, and the C-terminal domain of unknown function is colored in red. It is predicted to consist of ~5 contiguous α -helices, in an extended conformation potentially involved in protein-protein interactions.



5.3 Expression and purification of SsoCas3'

The gene encoding for Cas3' (Sso1440) was amplified by PCR from *S. solfataricus* genomic DNA cloned into the Gateway pDEST14 vector to enable expression of the recombinant protein with an N-terminal hexahistidine tag. The sequence encoding for the first 14 residues from the annotated N-terminus was omitted from the amplified gene as multiple sequence alignments indicated that this part was not conserved and was probably a misannotation event. The construct was sequenced and expression was carried out in *E. coli* C43 (DE3) cells. The recombinant protein was purified to homogeneity by nickel-chelating and size-exclusion chromatography, as described in Materials and Methods. The protein eluted as a monomer from a calibrated analytical gel-filtration column (Superose 6 HR 10/30, Amersham Biosciences). The poly-histidine tag was cleaved by overnight incubation with the tobacco etch virus (TEV) protease. All steps were carried out on ice due to protein degradation at elevated temperatures. Maldi-TOF mass spectrometry was used to verify the integrity of the protein. The apparent molecular weight of the his-tagged and native protein as observed on SDS-PAGE were in agreement with the calculated molecular weights of 59.406 kDa and 56.363 kDa respectively. The final purification step can be seen in figure 5.8.

Expression levels were high but degradation of the N-terminus occurred under all purification conditions tested, reducing the amount of active protein in the final sample. This is particularly common in multi-domain proteins such as helicases, as the flexible interdomain loops tend to be exposed and prone to protease degradation. Typical yields ranged between 0.45 - 1.4 mg/L of culture.

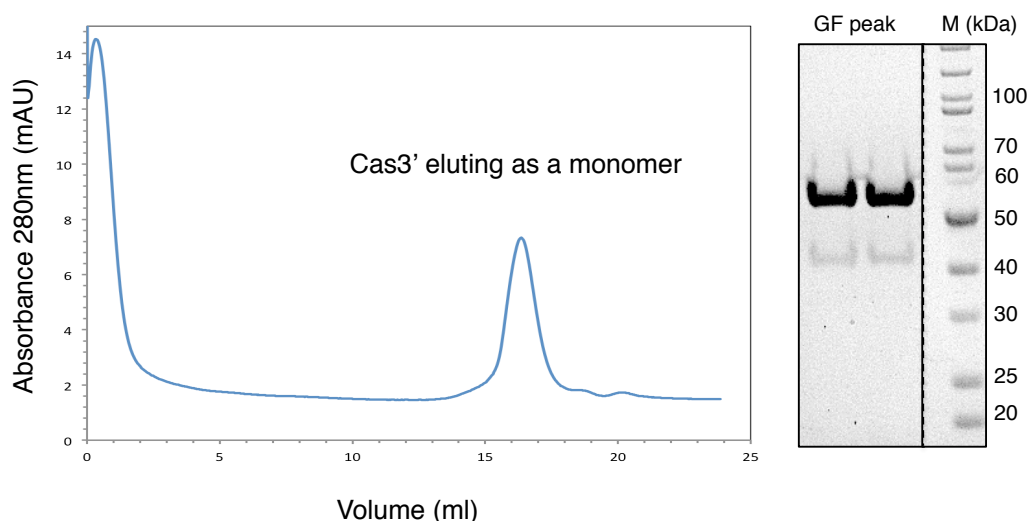


Figure 5.8: Purification of recombinant SsoCas3'

Chromatogram of SsoCas3' ran on an analytical Superose 6 HR 10/30 on the left, where Cas3' elutes as a monomer with an apparent MW of 35 kDa according to the standard calibration curve. It is unlikely that this peak represents just degradation products, SDS-PAGE analysis of the peak fractions (shown on the right) revealed that the peak composed predominantly of the intact protein (intense band at ~56 kDa). The fainter band at ~42 kDa was identified as N-terminal degradation product and was present in all protein preparations. It is unknown why the protein's apparent molecular weight is smaller than its actual MW, but its multi-domain and flexible nature could result in a different behaviour than globular proteins.

5.4 Site-directed mutagenesis of SsoCas3' (Sso1440)

The ATP-binding Walker A motif in SsoCas3' was identified by multiple sequence alignments, and consists of the amino acid sequence APTGYGKT (residues 40-47). The conserved lysine in position 46 which is essential for nucleotide binding was mutated to an alanine residue (referred to as K46) by site-directed mutagenesis using the QuikChange II Site-Directed Mutagenesis Kit (Stratagene). The gene was sequenced to confirm the mutation and was expressed and purified as described for the wild-type protein. The amino acid substitution was verified by ESI-TOF mass spectrometry (figure 5.9). The SsoCas3' K46A mutant would serve as a negative control for ATPase and helicase activity of the WT protein.

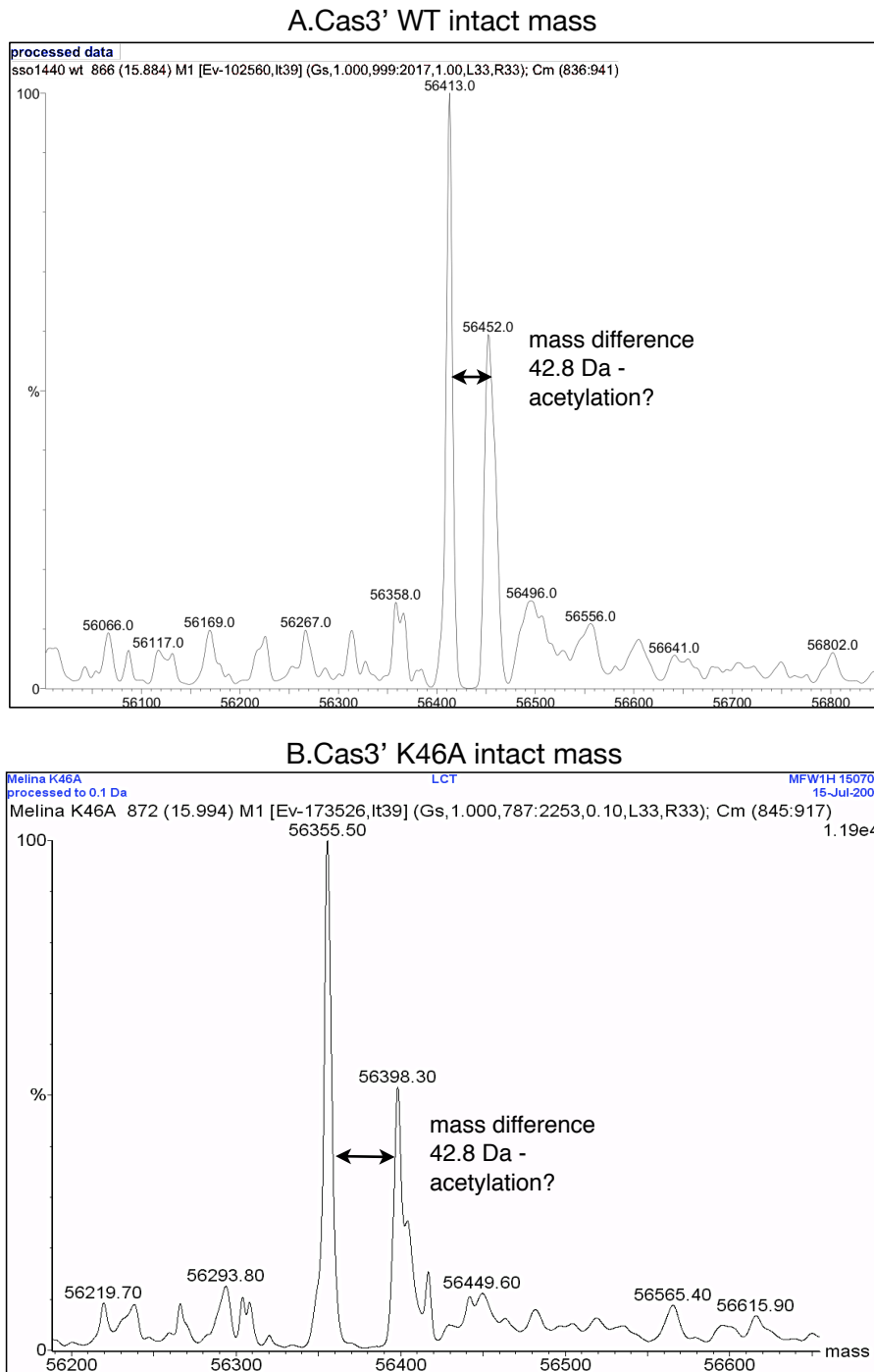


Figure 5.9: ESI-TOF mass spectrometry of SsoCas3' WT and K46A

Intact molecular weights of proteins determined by ESI-TOF mass spectrometry to confirm the site mutation. (A) Mass spectrum for the WT protein. The major peak corresponds to a molecular weight of 56413 Da (data processed to 0.1 Da). (B) Mass spectrum for Cas3'-K46A. The major peak corresponds to a molecular weight of 56355.5 Da (data processed to 0.1 Da). The mass difference of 57.5 Da is in agreement with a replacement of a lysine to an alanine. A second protein species was observed in both samples, with a consistent mass difference of 42.8 Da. It is suggested that this represents an acetylation event.

5.5 ATPase activity of SsoCas3'

NTP hydrolysis is the common mechanism all helicases use to generate energy to translocate along single strand nucleic acids and catalyse strand separation. The hydrolytic reaction can be stimulated by the presence of nucleic acids, single or double stranded. Binding of the appropriate nucleic acid induces conformational changes that enable efficient NTP binding and hydrolysis. SsoCas3' contains all the characteristic helicase motifs (I, II, VI) to catalyse NTP hydrolysis, so the ATPase activity of the protein was investigated by the malachite green colorimetric phosphate assay, as described in chapter 2. Wild-type and SsoCas3' K46A (Walker A mutant) protein (2 μ M) was incubated in the presence or absence of ss- or ds DNA, RNA or DNA/RNA heteroduplexes at temperatures ranging from 37°C - 65°C, in the presence of 1 mM MgCl₂. All nucleic acid substrates used were annealed or single-strand oligonucleotides 25-40 nt in length (CRISPR-related substrates, tables 5.1, 5.2). Samples were taken across a 20 min time course and free phosphate levels were visualised with malachite green, a reagent producing a colour change upon interaction with free phosphate that can be monitored by absorbance at 650 nm. The intensity of absorbance is analogous to the levels of free phosphate in the sample, providing a qualitative assay to measure the rate of ATP hydrolysis. Results are summarised in figure 5.10.

Levels of ATP hydrolysis were significantly higher in the presence of ssDNA, compared to ssRNA, dsDNA or dsRNA. Minimal levels of free phosphate were detected in the absence of nucleic acid, due to the fact that enzyme conformational flexibility may lead to a basal level of ATP hydrolysis (Soulтанas & Wigley 2000). The mutation of the conserved lysine to alanine in the Walker A motif of SsoCas3' would render it incapable of hydrolysing ATP as this lysine interacts with the β -phosphate and acts to stabilise the transition state of the hydrolytic reaction (Tuteja and Tuteja, 2004). Indeed, the SsoCas3' K46A mutant exhibited basal background levels of ATP hydrolysis. This control also confirms that the ATPase activity observed is attributed to SsoCas3. Reaction rates, although extremely low compared to other helicases, were almost 5-fold higher when ssDNA was present, indicating an ssDNA-stimulated ATPase activity for SsoCas3'.

The effect of temperature on the ATPase reaction was also investigated, in order to determine the optimum temperature range for protein function. Reactions were carried out at 37°C, 45°C, 55°C and 65°C under identical conditions for 30 min with 1 μ M protein and in the presence of ssDNA. Control reactions with only ssDNA were run in parallel to obtain the background levels of spontaneous ATP hydrolysis at different temperatures, which revealed a basal level of 6.24 pmoles phosphate.min⁻¹. The levels of hydrolysed ATP increased with the temperature rise, with the optimum activity obtained at 55°C after which the ATP hydrolysis rate dropped. A high

temperature optimum is expected from an enzyme by a thermophilic organism such as *S. solfataricus*. Optimum growth for this archaeon is observed at 80°C, therefore it would be expected that the enzyme would be more active approaching this temperature. The fact that SsoCas3' exhibits highest activity at 55°C could be related to the fact that the context of Cas3' function in vivo differs greatly from the minimum experimental set up presented here, as it is predicted to interact tightly with an HD-domain nuclease and potentially a CASCADE-like complex.

Considering that the natural Cas3' substrates would most likely include an R-loop, we also monitored the ATPase activity in the presence of a 25-base pair ds RNA-DNA heteroduplex. Reaction rates were comparable to the rates obtained in the presence of dsDNA and dsRNA, indicating that this type of substrate does not stimulate ATP hydrolysis. We can therefore infer that SsoCas3' exhibits an ssDNA-dependent ATPase activity, in agreement with the results reported for the *Streptococcus thermophilus* Cas3 (Sinkunas *et al.* 2011). However, the reaction rates for SthCas3 were reasonably higher than SsoCas3, reflecting the processive helicase activity for this DExH-box protein. One explanation could be suggested considering the differences in ATP hydrolysis modes between DExH-box and DEAD-box proteins as outlined in section 5.1.2. Even though SsoCas3' is a DExD-box family protein, it lacks the additional domains of SthCas3 that potentially needed to maintain a high rate of ATP hydrolysis, therefore mechanistically resembling DEAD-box proteins' mode of action.

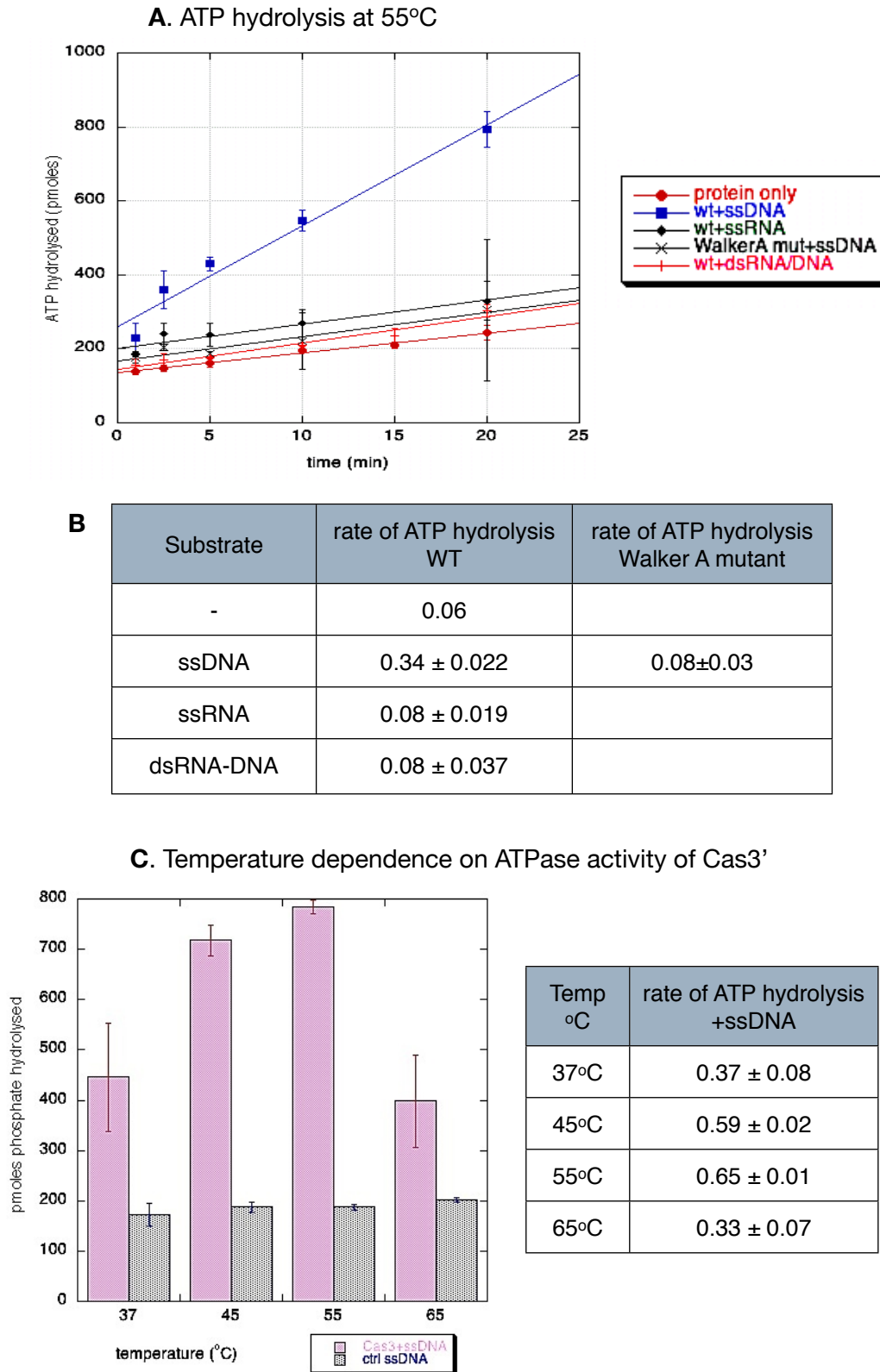


Figure 5.10: ATPase activity of WT and mutant SsoCas3'

(A) Course of ATP hydrolysis by SsoCas3' in the presence of ssDNA, ssRNA and RNA-DNA hybrids at 55°C. A linear curve fit was applied to the data points. ATP hydrolysis in the presence of dsDNA and dsRNA was at the same levels as for dsRNA-DNA (not shown). ATPase activity although very low is clearly stimulated in the presence of ssDNA. Reaction rates in table (B) in pmoles phosphate. pmoles SsoCas3' ⁻¹.min⁻¹. (C) ATPase activity of SsoCas3' in the presence of ssDNA at different temperatures illustrated by pink blocks. Background levels of ATP hydrolysis in the presence solely of ssDNA are presented in blue blocks.

5.6 Helicase activity and substrate preference of SsoCas3'

The helicase motifs present in SsoCas3' indicate that it may have a duplex unwinding ability. In the context of the CRISPR system, this could involve either dsDNA, dsRNA or DNA-RNA heteroduplexes. To investigate the helicase activity of SsoCas3' along with its respective substrate specificity and directionality of unwinding, 5' fluorescein-labelled DNA and RNA oligonucleotides corresponding to the 25 nt CRISPR repeat sequence of locus B with appropriate 3' or 5' 15-U or 15-T overhangs (table 5.1) were purified and annealed to produce the double-strand substrates seen in table 5.2. Each substrate consists of a 25-base pair duplex region and a 15 nt poly-uracil or poly-thymine 3' or 5' extension (3' overhang or 5' overhang). The sequences of the single-stranded synthetic oligonucleotides used to generate the duplex substrates can be seen in table 5.1. A graphic representation of the basic structure of the double-stranded substrates can be seen in table 5.2.

	ss oligonucleotides used to make the ds substrates labeled with fluorescein (*) or [γ -P ³²]ATP(**)	sequence 5' to 3'
CRISPR-related substrate set #1	ssDNA	*CTTTCAATTCCCTTTTGGGATTAATC
	ssRNA	*CUUUCAAUCCUUUUUGGGAUUAAUC
	complement sequence. 15-U or 15-T were added to the 5' or 3' end to make overhang ds substrates	GATTAATCCCAAAGGAATTGAAAG (also in RNA form)
non CRISPR-related substrate set #2	ssDNA	**GCTCCTAGGTCCTTCGTGGCATCTG
	ssRNA	**GCUCCUAGGUCCUUCGUGGCAUCUG
	complement sequence \pm 15-U or 15-T to make ds substrates	CGAGGATCCAGGAAGCACCGTAGAC (also in RNA form)
crRNA-protospacer substrate set #3	crRNA-A1	**AUUGAAAGGAACUAGCUUAUAGUUUAGAAG AAAACAAACAAAUAU GAUUAUCCCAAAA
	tA1f +PAM	**TAATACGACTCACTATAGGGT ATTATTTGTTTGTTCCTTCTAAACTATAAGC TAGTTC TGGAGAGAAGGTG

Table 5.1 Oligonucleotides used to generate substrates used in chapter 5

In substrate set #3, red fonts indicate the complementary spacer regions and blue the 5' psitag. Single asterisks indicate the fluorescein label, double asterisks the radiolabel.


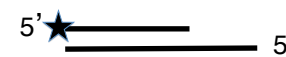




Substrates	structure
3' overhang (3'oh) dsDNA or dsRNA	3' 
5' oh dsDNA or dsRNA	
3' oh RNA-DNA	3' 
5' oh RNA-DNA	
5' oh DNA-RNA	
blunt RNA-DNA hybrid	

Table 5.2: Structures of ds substrates used for helicase assays

A star represents the 5' fluorescein-labeled or radiolabeled strand. RNA is indicated by a dashed line, DNA by a solid line.

Initial assays revealed that SsoCas3' was unable to unwind dsDNA or dsRNA substrates carrying either a 5' or 3' single-strand overhang region, but that it exhibited ATP-dependent activity against DNA-RNA heteroduplexes (figure 5.11). Assays were carried out at 37°C due to instability of the RNA-DNA heteroduplex at elevated temperatures, even though unwinding was faster at 45°C, therefore it is possible that the temperature optimum for protein activity is even higher than that (figure 5.12). The protein was pre-incubated with the respective substrate for 1 min at 37°C and the assay was initiated with the addition of 1 mM ATP. Substrate unwinding was followed over a 5 - 20 min time course, depending on the experiment, and products were analysed on native 12% polyacrylamide:TBE gels, as described in Materials and Methods. Appropriate controls to ensure the substrate stability, indicate the size of the unwound product and the absence of background activity without ATP were carried out in each experiment. About 5% of the substrate was unwound even at 37°C in the absence of protein. In the negative controls however, where the substrate is incubated with the protein in the absence of ATP, single strand unwound product was not observed, indicating that SsoCas3' binds to the duplex and enhances its stability.

SsoCas3' was able to successfully unwind both 3' and 5' overhang DNA-RNA heteroduplex substrates in the presence of ATP, regardless of whether the overhang region was DNA or RNA, suggesting a bidirectional helicase activity. The rate of unwinding increased proportional to the protein concentration, supporting that the effect was not caused by substrate melting (figure 5.13). The protein's performance was high but inconsistent with different protein batches or reaction conditions making it difficult to calculate reliable unwinding rates, even though unwinding of

heteroduplexes was always observed in an ATP/MgCl₂-dependent manner. Since the rate of unwinding of both 3' and 5' overhang substrates was comparable, blunt DNA-RNA heteroduplex substrates were prepared, consisting only of a 25-base pair duplex region, with sequence corresponding to the CRISPR repeat. SsoCas3' was able to efficiently unwind the blunt heteroduplex with roughly the same rate as the 3' and 5' overhang heteroduplexes (figure 5.14A). Unwinding was not observed if ATP was replaced by non-hydrolysable ATP analogues, such as AMP-PNP (5'-adenylyl-β, γ-imidodiphosphate) and ATPγS (adenosine 5'-O-(3-thio) triphosphate) (figure 5.14C). The Walker A mutant Sso1440-K46A which disrupts the ATP-binding domain of the protein was unable to function as a helicase confirming the existence of an energy-driven catalytic unwinding mechanism (figure 5.14A). From this set of results it is not clear whether the protein translocates on the RNA or DNA strand of the heteroduplex, although the ssDNA-stimulated ATPase activity described in the previous paragraph suggests that it could be the DNA strand.

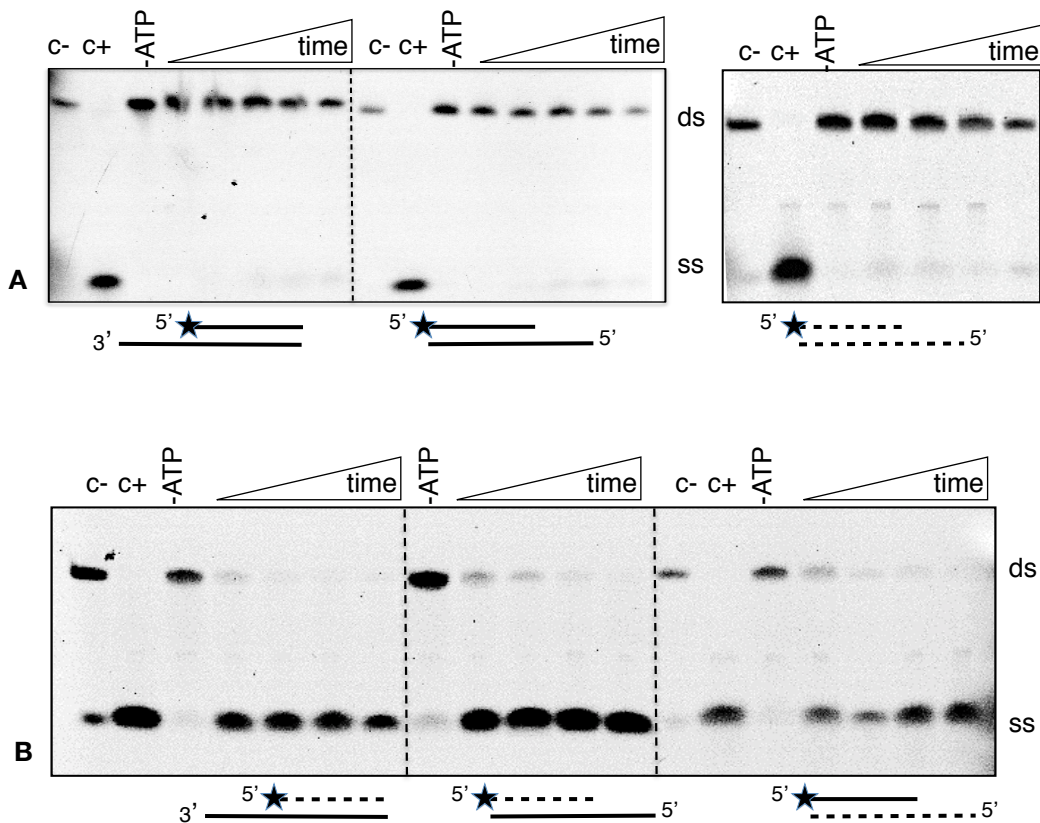


Figure 5.11: Helicase activity of SsoCas3' on CRISPR substrates

Helicase activity was monitored as described in the text. The substrates used in each assay are depicted under each panel, corresponding to the structures as outlined in tables 5.1 and 5.2. The controls performed for each substrate are marked as: c+, size marker for the unwound product, boiled at 95°C for 2min; c-, end-point reaction without protein; -ATP, end-point reaction without ATP/ MgCl₂. (A) Left panel, reaction with dsDNA substrates and 250nM SsoCas3. Time course: 2', 5', 10', 20, 45'. Right panel, dsRNA substrate. Time course: 1', 5', 15'. (B) Reactions with 200nM SsoCas3' and 3'/5' overhang RNA-DNA substrates. Time course: 30", 1', 2', 5'.

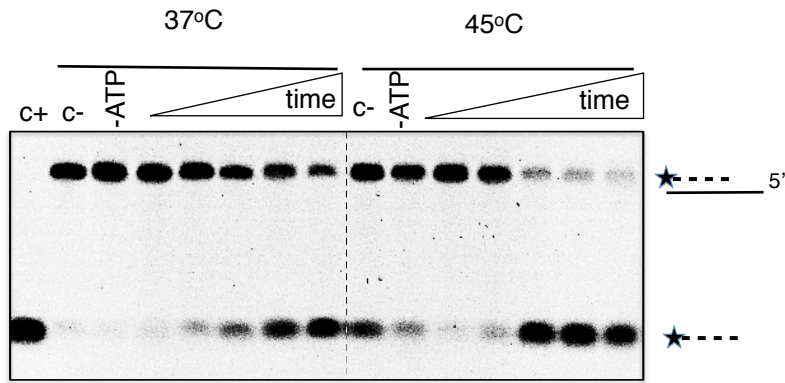


Figure 5.12: Temperature dependance on the helicase activity of SsoCas3' Protein concentration was 250 nM and samples were taken at 1', 2', 5', 10', 20.

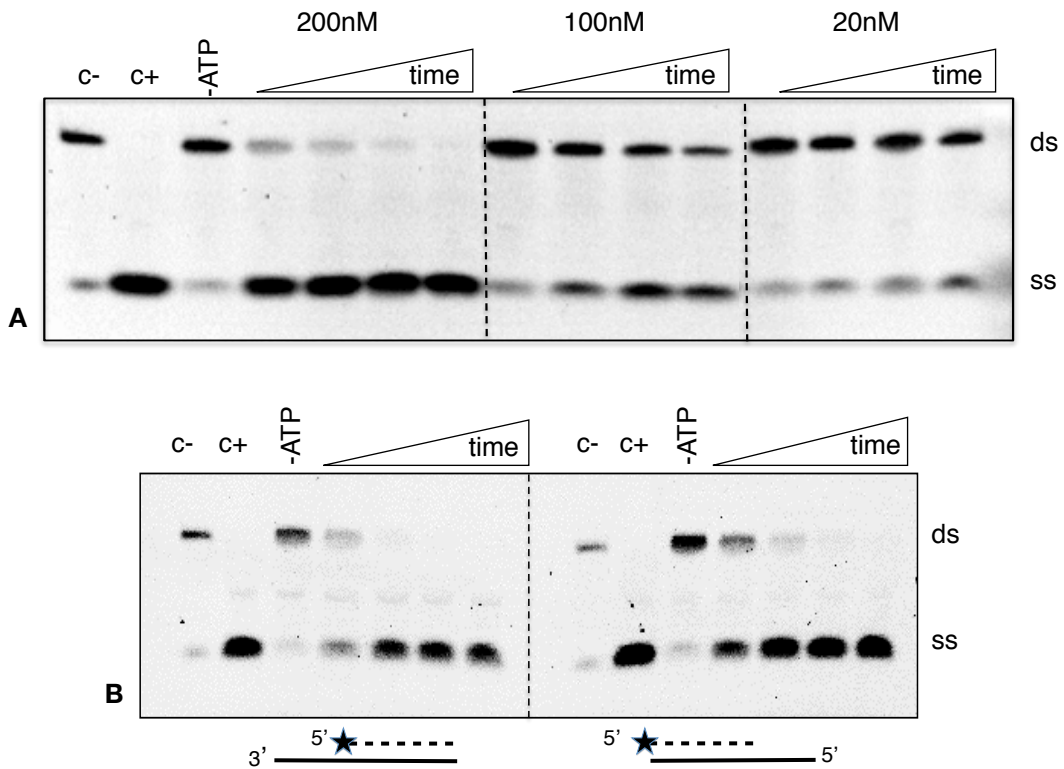


Figure 5.13: Helicase activity of SsoCas3' is dependent on protein concentration

(A) Monitoring the influence of protein concentration (200 nM, 100 nM, 20 nM) at constant substrate concentration (5' overhang RNA-DNA, 20 nM) on the helicase activity of SsoCas3. Time course: 15", 30", 1', 2'. Almost 100% unwinding is observed at 10-fold excess of protein over substrate (200 nM), as opposed to very little unwinding at a 1:1 ratio. Reasonable course of unwinding is observed with 100 nM protein (5:1 ratio), suitable to compare efficiency of unwinding of different substrates. (B) Helicase activity on 3' and 5' overhang RNA-DNA substrates with 100 nM protein. Time points: 30", 1', 2', 5'. Inconsistencies in the protein's efficiency were observed among different purification batches or reaction conditions, leading to differences in unwinding rates such as between panels A and B.

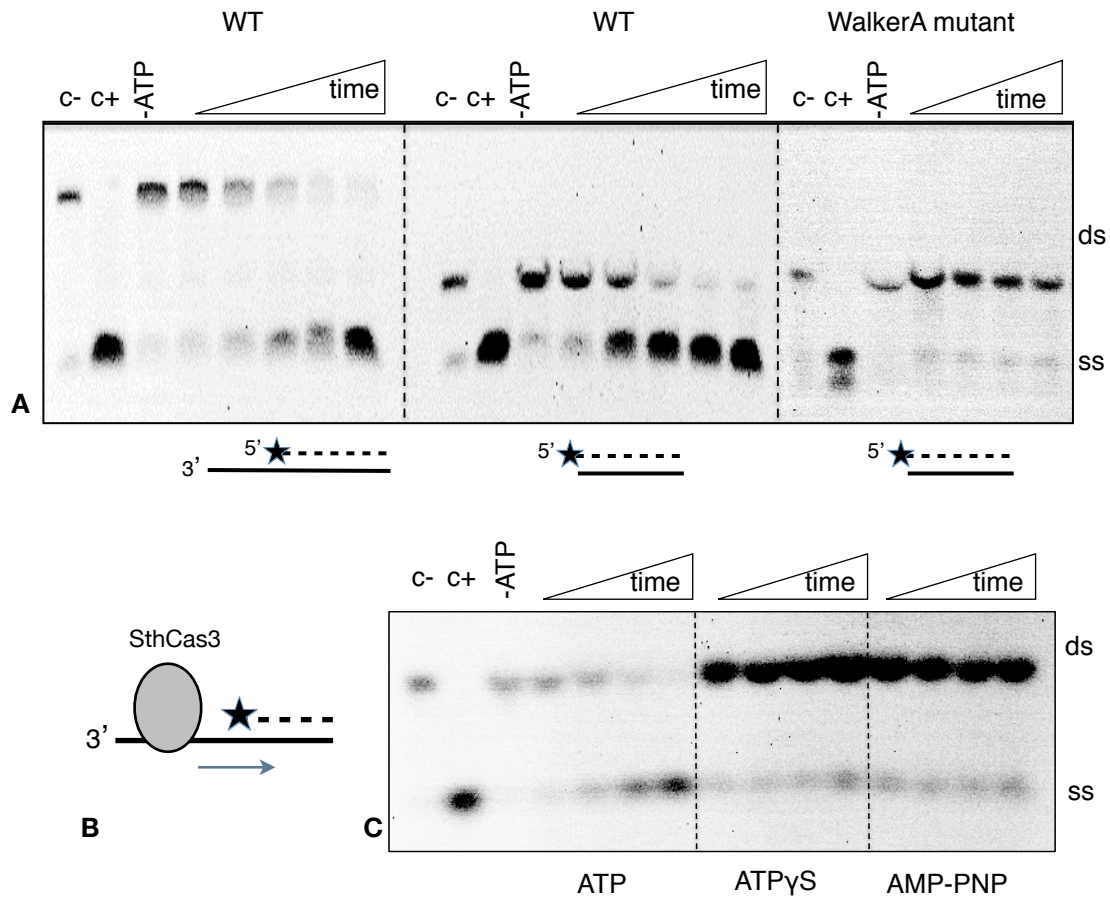


Figure 5.14: Helicase activity of SsoCas3' on blunt RNA-DNA heteroduplexes

(A) Comparison of the WT and WalkerA mutant activity on 3' overhang and blunt RNA-DNA duplex substrates. Protein concentrations are 50 nM (WT) and 100 nM (WalkerA mutant). Time course: 30" (only in WT), 2', 5', 10', 20'. (B) Cartoon representation of the substrate preference and polarity demonstrated by Cas3 from *S. thermophilus* (Sinkunas *et al.* 2011). The displaced (labeled) strand could be either RNA or DNA. (C) Helicase activity is not supported by non-hydrolysable ATP analogues. Protein concentration: 100 nM; substrate 20 nM blunt heteroduplex; time course: 30", 2', 5', 10'.

The ability of SsoCas3' to unwind blunt heteroduplexes is in contrast to the 3' - 5' directionality reported for the *S. thermophilus* Cas3 (subtype I-E) (Sinkunas *et al.* 2011). The Cas3 protein in this subtype however, is fused to the HD phosphohydrolase predicted to perform the degradation of the invader DNA (Brouns *et al.* 2008). Given the partial substrate melting in the control samples, it could be possible that the unwinding activity of Cas3 on blunt substrates results from single stranded DNA or RNA exposed as a result of thermal fraying, which would enable the loading of the enzyme onto the substrate and subsequent translocation and duplex disruption. In the absence of a crystal structure of the two orthologs, it remains unknown whether in the fusion protein the HD nuclease domain affects the function of the helicase domain by inducing conformational changes, thereby resulting in a different type of substrate recognition and mode of action. Nevertheless, potential cross-talk between the two domains has been suggested before. Howard *et al.* (2011) observed that mutation of

the HD motif inhibited R-loop formation by Cas3 in *E. coli* (subtype I-E), but this effect was reversed in the double Walker A - HD mutant.

In order to further investigate this matter and the substrate requirements of SsoCas3', two additional sets of substrates were prepared. The first set (set #2 in table 5.1) consisted of a series of 25 nt RNA - DNA heteroduplexes with non-CRISPR related sequence, either blunt or with 3'/5' RNA or DNA overhangs. The oligonucleotides were purified, 5' end labeled with [γ - 32 P] ATP and annealed to produce the double strand substrates. Radiolabeling was used instead of fluorescein-labeling to investigate whether the comparatively large size of fluorescein was altering the protein's behavior. Curiously, under identical assay conditions SsoCas3' was unable to unwind any of the 5' overhang, 3' overhang or blunt heteroduplexes (figure 5.15). Altering the assay conditions (increasing or decreasing the buffer concentration of K-glutamate, increasing the protein concentration) did not stimulate the activity. Only on a single occasion unwinding of a 5' RNA overhang heteroduplex was observed, but this result was not repeatable. Apart from experimental error (failure to determine the optimum conditions for activity), an explanation would be the requirement for certain structural features or repeat-related sequences (one of the 5' or 3' repeat-derived handles in mature crRNA sequences) in the duplex substrate to be recognized and bound by SsoCas3' in order to perform the duplex displacement. It is uncommon for helicases to exhibit sequence specificity, unless the SsoCas3' C-terminal domain of unknown yet function plays a role.

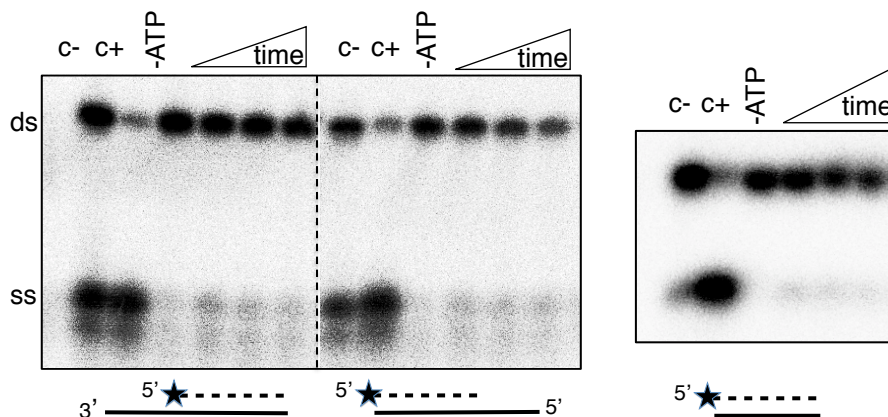


Figure 5.15: SsoCas3' is unable to unwind radiolabeled non-CRISPR related substrates
 Reactions were carried out at 37°C due to substrate instability, at a protein concentration of 200 nM. The same results were obtained with RNA overhangs and a DNA labeled strand. Data shown are representative of three repeat experiments. Time course: 2', 5', 10'.

In order to investigate this possibility, a third set of RNA-DNA heteroduplex substrates was generated (#3 table 5.1) that closely resembled the natural substrate *in vivo*. A single strand crRNA oligonucleotide (crRNA-A1) was annealed to a ssDNA sequence carrying a matching protospacer (tA1f, table 4.3), such that the produced heteroduplex consists of a 38-base pair duplex region flanked by unpaired forked strands (figure 4.18 D). This minimal substrate corresponds to the *in vivo* state upon recognition of the DNA invader by the Cas effector proteins - crRNA complex and subsequent binding, in some cases by displacement of the non-complementary strand and formation of an R-loop (Jore *et al.* 2011; Howard *et al.* 2011). SsoCas3' was unable to unwind this substrate over a 60 min time course at 37°C or 45°C (higher temperatures were avoided due to substrate instability) (figure 5.16). This was unexpected, as this substrate contains most of the characteristics of the physiological substrate, with the exception of the R-loop context. Possible explanations for this will be presented in the discussion section.

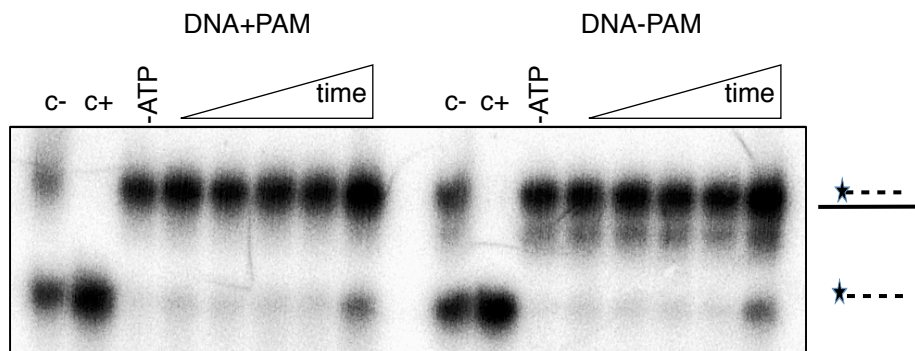


Figure 5.16: SsoCas3' is unable to unwind “crRNA-protospacer”-like substrates

Protein concentration was increased to 500nM over ~25nM substrate. Time course: 1', 2', 5', 10', 30'. The small quantity of single-stranded species in the last time point is unlikely to reflect protein activity, but is probably a result of substrate instability and well overloading. Two protospacer constructs were used, with and without an appropriate PAM sequence (see previous chapter), to test the hypothesis by Sinkunas *et al.* on potential PAM recognition at this stage. No unwinding was observed with either substrate.

5.6.1 SsoCas3' is not able to process long duplex regions

To assess the processivity of SsoCas3' on longer duplex regions that the minimal 25 nt substrate used in the assays described before with the first set of substrates, a 151-base pair RNA-DNA heteroduplex with a 20 nt DNA overhang on the 3' end was generated. This substrate consisted of the *in vitro* transcribed initial two repeat-spacer units of CRISPR locus A described in chapter 4, and its complementary DNA. SsoCas3' was incubated with this substrate for 60 min at 37°C in the presence of ATP, but no unwinding was observed (figure 5.17). Considering the proposed role

for this protein in the interference stage of CRISPR functioning, where the natural substrate for this protein would be the heteroduplex formed by the spacer region of the crRNA with the DNA target, this finding is not unexpected, although the Cas3 orthologue from *E. coli* and *S. thermophilus* were reported to unwind even longer duplexes between 70 - 1000 bp (Sinkunas *et al.* 2011; Howard *et al.* 2011).

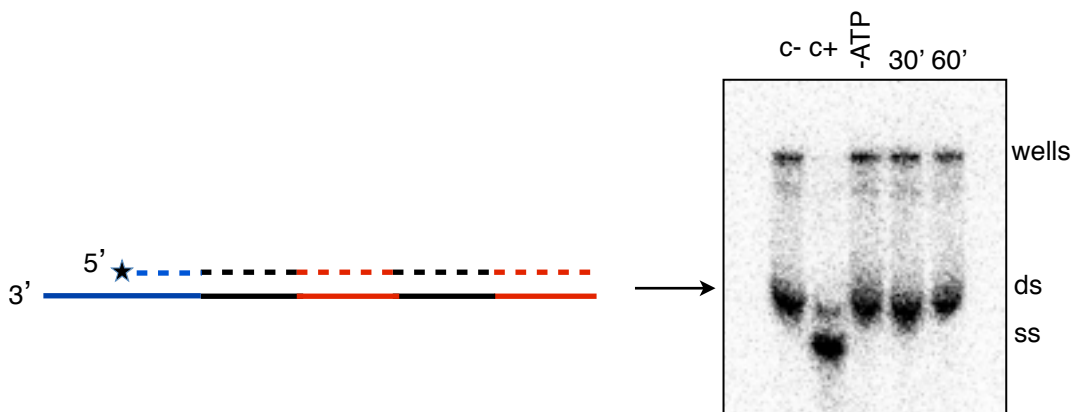


Figure 5.17: SsoCas3' is unable to unwind long duplex substrates.

The duplex substrate consists of an *in vitro* transcribed CRISPR RNA transcript of two repeat-spacer units, annealed to a complementary DNA with a 3' overhang. Repeats are in black, spacers in red, the part of the leader sequence and the T7 promoter on the 5' end is in blue. RNA transcript is indicated by a dashed line, complementary DNA sequences are colored with the same color scheme in a solid line. The radiolabeled phosphate is indicated with a star.

5.7 Nucleic acid binding by SsoCas3'

Electrophoretic mobility shift assays were carried out to investigate the nucleic acid binding characteristics of SsoCas3'. Given the fact that ssDNA stimulates ATP hydrolysis by SsoCas3', it would be expected that the protein exhibits a higher affinity for ssDNA. Fluorescence anisotropy would be a significantly more sensitive and accurate indicator of nucleic acid binding, but the large amounts of protein required were unobtainable for SsoCas3'.

Increasing concentrations of SsoCas3' were incubated with 20 nM of single strand DNA or RNA oligonucleotides corresponding to CRISPR locus B repeat sequences under the same conditions used to test helicase activity (20 mM MES pH 6.5, 100 mM potassium glutamate, 1 mM MgCl₂, 1 mM ATP, 0.1 mg/ml BSA, incubation at 37°C for 30 min). Both oligonucleotides appeared to be effectively shifted, indicating complex formation with SsoCas3' with an apparent dissociation constant of 5 μM (figure 5.18). This high K_d value could be attributed to the general non-specific affinity for nucleic acids this protein is expected to exhibit, and does not reveal any information about the specific characteristics of substrate recognition. Unfortunately, attempts to optimise binding failed, either due to transient interactions,

protein degradation in the sample, protein aggregation or the assay limitations explained elsewhere.

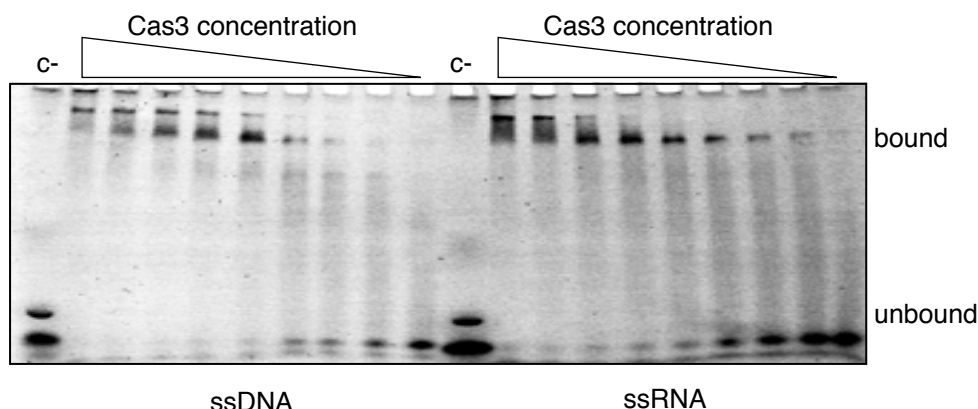


Figure 5.18: Nucleic acid binding by SsoCas3'

SsoCas3' is able to bind both ssDNA and ssRNA (sequences in table 5.1). Concentration range (in μM): 10, 8, 7, 6, 5, 4, 3, 2, 1. At high protein concentrations two bands are observed representing complexes with increased protein:NA stoichiometry.

5.8 Strand annealing and strand exchange activity of SsoCas3'

A number of helicase families involved in DNA recombination and repair have been reported to exhibit strand annealing and strand exchange ability, such as the prokaryotic RecA, RadA and RecQ and the eukaryotic Rad51 (Cox and Lehman, 1982; Seitz *et al.* 1998; Seitz and Kowalczykowski, 2006). Taking into account the proposed role for Cas3 during the interference stage as outlined by Jore *et al.* (2010), whereby R-loop formation promoted by CASCADE would be followed by degradation of the invader DNA by Cas3 in *E. coli*, it would be reasonable to consider a potential annealing/strand exchange activity for SsoCas3'. Indeed, Cas3 from *E. coli* (which is fused to an HD-nuclease domain, existing as a separate protein in *S. solfataricus*) has been shown to promote annealing of two complementary DNA strands, and most interestingly of ssRNA to complementary duplex DNA in the form of an uncut plasmid, resulting in the formation of an R-loop (Howard *et al.* 2011). This activity was shown to proceed in an ATP-independent fashion, but required magnesium as a co-factor and, curiously, an active HD-nuclease motif. Addition of ATP resulted in the reverse activity of unwinding the R-loops, raising questions about the co-existence and regulation of these antagonistic activities *in vivo* (Howard *et al.* 2011).

We investigated the ability of SsoCas3' to catalyse duplex formation using combinations of purified 25 nt 5' fluorescein-labeled single strand DNA or RNA oligonucleotides (table 5.1). Assays were carried out at 37°C over a 20 min time course in the presence of 1mM MgCl₂ and products were analysed on 12% native polyacrylamide gels. The reason such low temperatures were used even though the

enzyme has proven to exhibit optimum activity at 55°C was the instability of the RNA-DNA heteroduplexes at elevated temperatures. Firstly, we observed efficient annealing of ssRNA to complementary ssDNA in the presence of SsoCas3', as opposed to no heteroduplex formation in the absence of the protein (figure 5.19).

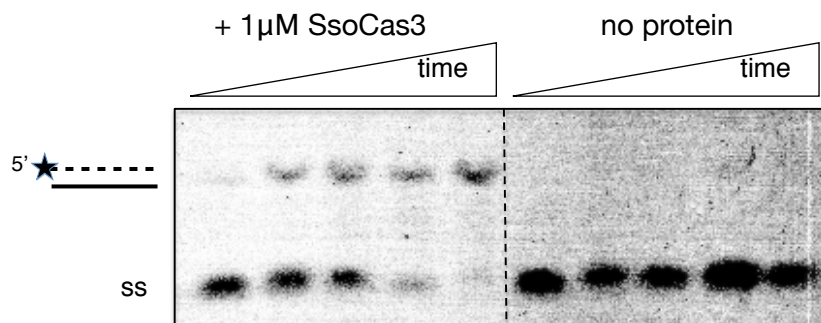


Figure 5.19: Strand annealing activity of SsoCas3'

Two complementary separately purified strands (RNA and DNA) were denatured at 90°C for 2 min and then incubated for a maximum of 20min at 37°C in the presence or absence of SsoCas3'. We can observe almost 80% annealing when SsoCas3' is present. Dashed line indicate non-contiguous lanes on the same gel.

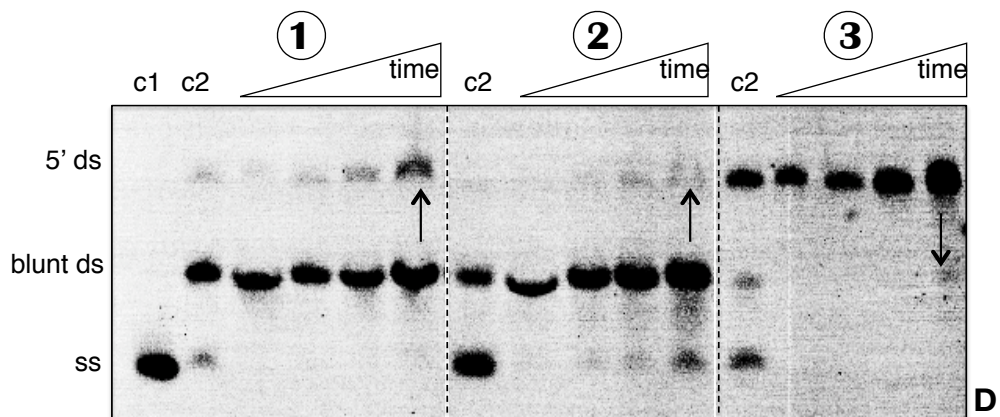
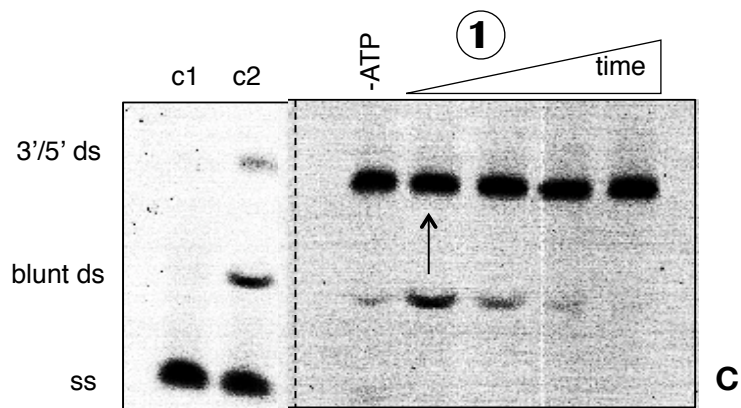
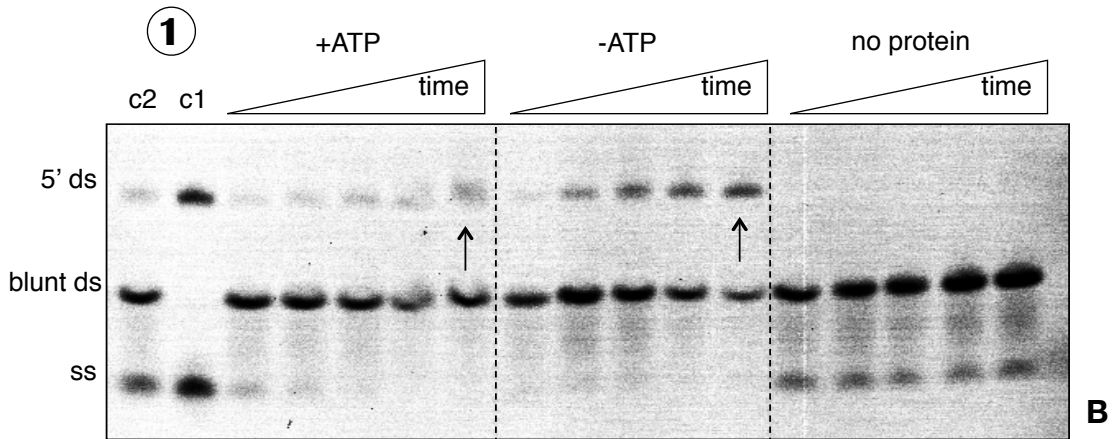
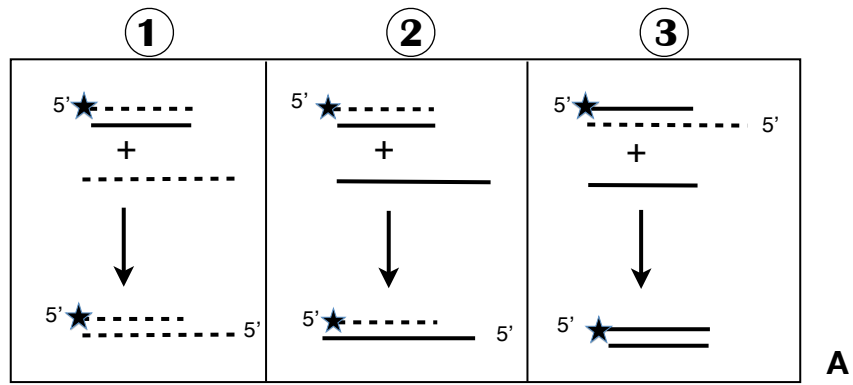
Secondly, SsoCas3' promoted the invasion of a single RNA strand into a complementary 25-base pair DNA-RNA heteroduplex (substrate), whereby it replaced the DNA strand and generate a dsRNA (product). The experimental setup is outlined in figure 5.20 A (reaction type 1). In this experiment, ~50% of the DNA strand in the labeled substrate was exchanged for the unlabelled longer RNA strand to produce the double stranded RNA product, as opposed to 0-5% in the absence of protein (figure 5.20 B). The fact that ATP hydrolysis is not required for the strand exchange reaction and the small size of the participating oligonucleotides indicates that the exchange is likely to take place due to passive binding of both the ss- and the ds- species, so that strand rearrangement would proceed until equilibrium is reached. If that was the case, in the presence of ATP the result would have been the same, since the substrate heteroduplex would have been unwound actively by the protein (these substrates were shown to be effectively unwound during helicase experiments) and formation of the more stable dsRNA (which the protein cannot unwind) would have been favoured by passive binding. A possible explanation could be that at the low assay temperatures (30°C, due to substrate instability) the helicase activity of SsoCas3' was not efficient. The reaction was repeated at 37°C, where indeed formation of the product dsRNA was almost 100% in the presence or absence of ATP (figure 5.20 C). The RNA-DNA heteroduplex substrate is more unstable at this temperature (control 2), which promotes the strand rearrangements and formation of the stable dsRNA in the absence of ATP. This is however still a protein-dependent procedure, as can be seen from the minimal formation of dsRNA in the control (c2).

In order to clarify the requirements of SsoCas3' in terms of the participating nucleic acid species, we investigated whether it could promote annealing of a ssDNA to a complementary RNA-DNA hybrid to displace either strand (reactions 2 and 3, figure 5.19 panel A), in the absence of ATP and at 30°C (figure 5.20 D). Strand exchange was not supported in either case, confirming the requirement for the invading strand to be RNA. Additional experiments and controls are needed to clarify this activity and determine whether it is an active reaction mechanism, reminiscent of the *E. coli* Cas3, or a product of passive equilibrium binding of the nucleic acid species for which the protein exhibits affinity. Further experiments should include potential magnesium-dependency as well as replacement of SsoCas3' with other proteins with/without a general affinity for nucleic acids.

An RNA-DNA heteroduplex or a dsRNA most likely do not represent physiologically relevant substrates/products for the Cas3 protein family, especially in the light of recent studies elucidating part of its activity in type I-E systems (Jore *et al.* 2011; Sinkunas *et al.* 2011; Howard *et al.* 2011). In retrospect, a more appropriate experimental setup should have involved invasion of a ssRNA into a dsDNA substrate, mimicking the displacement of the DNA strand during target recognition by the crRNA-loaded CASCADE-like complexes (Jore *et al.* 2011). The experiments presented in this paragraph were carried out prior to the publication of any of the aforementioned studies, and were halted due to increasing problems with enzyme production and stability and time constraints. Nevertheless, some useful observations can be made. In summary, these results suggest that SsoCas3' effectively stimulates ATP-independent strand annealing to generate RNA-DNA heteroduplexes, and the invasion of an RNA strand into a complementary heteroduplex. Combining these two observations, and in line with the emerged reports on the activity exhibited for *E. coli* Cas3, it is reasonable to speculate that SsoCas3' should be able to promote invasion of an RNA strand into a dsDNA duplex to form an RNA-DNA heteroduplex, or R-loop.

Figure 5.20: Strand exchange activity of SsoCas3' (following page)

(A) Outline of the types of exchange reactions. Substrates are the CRISPR-related set #1 used for helicase assays (tables 5.1, 5.2). Region of complementarity is 25 bp. (B) Exchange reaction 1 as outlined in (A). Influence of ATP. Protein concentration: 500nM. Reaction temperature: 30°C. Time course: 2', 5', 10', 20', 30'. Controls: c1, size marker; c2, end-point reaction A without protein. Almost 50% RNA strand exchange can be seen after 30 min incubation without ATP, in comparison to ~10% with ATP. No dsRNA is formed in the absence of protein. Assay was performed at a low temperature (30°C) due to substrate instability. Arrows indicate the product formation (C) Exchange reaction at a higher temperature of 37°C. Protein concentration: 500nM; Time range: 2'-20'. ATP was included unless indicated. Under these conditions the RNA-DNA hybrid substrate is more unstable and formation of dsRNA is promoted in a protein-dependent way. (D) Protein concentration: 1µM. Reaction temperature: 30°C. Time course: 5', 15', 30', 45'. We can observe product formation only in reaction 1, where the invading strand is RNA. No ATP was included in the reactions. Dashed lines indicate non-contiguous lanes on the same gel.



5.8.1 Initial attempt to investigate R-loop formation by SsoCas3'

To monitor whether SsoCas3' could indeed promote the formation of an R-loop, a pET151/D-TOPO plasmid was constructed harbouring a 28 nt protospacer complementary to spacer 1 in SsoCRISPR locus A. This would basepair to the central region of the 60 nt synthetic oligonucleotide ssRNA crRNA-A1 (table 4.3). Briefly, 5 μ M of protein were added to 50 ng of duplex DNA plasmid and 1 μ M of 5' [γ - 32 P] ATP-labeled crRNA-A1 and reactions were initiated with 1 mM of ATP/MgCl₂ mixture. Samples were incubated at 55°C for 1 hour (in 20 mM MES pH 6, 100 mM potassium glutamate and 0.5 mM DTT), deproteinised by proteinase K treatment at 37°C for 15 min and analysed on a 0.8% agarose gel. Products were visualised by EtBr staining and phosphorimaging. As the ability to generate R-loops is also a characteristic of the *E. coli* CASCADE (Jore *et al.* 2011; Howard *et al.* 2011), the reaction was repeated in the presence of the recombinant Csa2-Cas5a complex or native aCASCADE (provided by N. Lintner and M. Lawrence) as described in chapter 4.

Upon EtBr staining of the agarose gel we can observe that all three plasmid forms (supercoiled, linear, open circular) are present in all reaction and control samples (figure 5.21). However, a small amount of slow-migrating species appears caught in the wells in the reaction sample where both SsoCas3' and the recombinant Csa2-Cas5a complex are present (lane B_R, figure 5.21). Phosphorimaging of the gel reveals the location of the radiolabeled crRNA, which would only be visible if involved in a large size complex, as any free single strand RNA would have migrated off the gel and into the buffer. Closer examination of the phosphorimaging image reveals a single band corresponding to a radiolabeled species which, if the two gel images are overlaid, corresponds to the slow-migrating species caught in the wells in figure 5.20, lane B_R. These species could potentially represent R-loops or large ribonucleoprotein complexes resulting from insufficient proteinase K treatment. Radiolabeled species were not observed in reactions missing SsoCas3' or the recombinant Csa2-Cas5a complex.

Although these are only preliminary results, it could be a possibility that SsoCas6 and the core of the archaeal CASCADE (the Csa2-Cas5a complex) act cooperatively to promote target recognition via R-loop formation between the crRNA and the invading dsDNA. In this case, a question is raised about the absence of radiolabeled, slow-migrating species (potential R-loops) in the samples containing the native aCASCADE from *S. solfataricus*. Potential explanations include the very low protein concentration of the protein sample used and inactivation of the complex during long-term storage. Obviously it is unrealistic to make assumptions based on the minimal system assayed here since key system proteins are missing, such as the HD-nuclease and the additional components of the aCASCADE (e.g. Csa5, Cas8a2), therefore this aspect of Cas3 function in *S. solfataricus* remains to be elucidated.

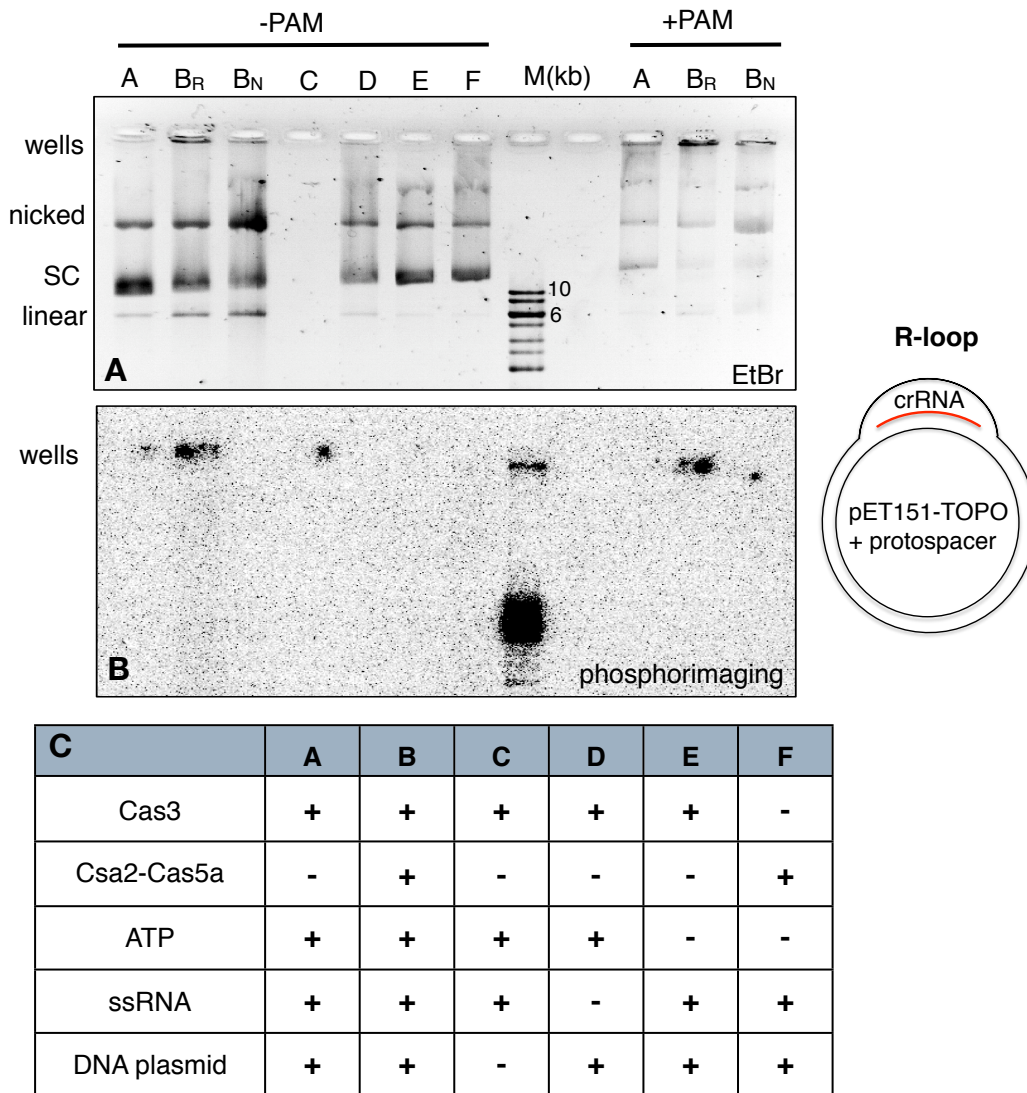


Figure 5.21: Potential R-loop formation by SsoCas3'.

UV visualisation (A) and phosphorimaging (B) of SsoCas3' reactions after agarose gel electrophoresis to detect radiolabeled RNA. Reaction components can be seen in the table C. Reactions were performed with recombinant SsoCsa2-Cas5a complex (B_R) or native aCASCADE (B_N) (see chapter 4). Signal due to the presence of radioactive RNA species is detected (B) in the marker lane (due to the radiolabeled DNA size ladder) and in lanes B_R. It is unknown why the linear species of plasmid DNA migrates faster than supercoiled plasmid. Restriction digestion of the plasmid (not shown) confirmed that the lower band observed in all DNA-containing lanes is indeed linear plasmid. An explanation could be the insufficient deproteinisation of the samples, resulting in reduced mobility for the protein-DNA complexes.

5.9 Discussion and future work

This chapter deals with the purification and initial biochemical characterisation of Cas3' from *Sulfolobus solfataricus*. The flexible nature of helicases renders them difficult to express in heterologous systems and study biochemically, therefore the information on the actual role of Cas3 within the CRISPR system has been limited. However, genetic studies and biochemical characterisation of the *E. coli* and *S.*

thermophilus Cas3 orthologues and type I-E systems have led to the hypothesis that Cas3 is involved in the terminal interference stage of CRISPR function (Brouns *et al.* 2008; Sinkunas *et al.* 2011; Howards *et al.* 2011; Cady and O'Toole, 2011; reviewed in Maraffini and Sontheimer, 2010; Al-Attar *et al.* 2011). The Cas3 orthologues characterised in these studies are helicase-nuclease fusions, in which both domains seem to be active with varying levels of efficiency. The degree to which the activities of the two domains have been seen to influence one another is not yet clear. In *E. coli* Cas3, mutations in the HD domain abolished R-loop formation, although a clear nuclease activity was not demonstrated under the conditions tested (Howard *et al.* 2011). In *S. thermophilus* Cas3, mutations in the HD motif had no effect on the protein's helicase activity or vice versa, although it was not examined whether this protein catalyses R-loop formation. Both orthologues exhibited processive 3' to 5' unwinding of dsDNA and DNA-RNA duplexes, translocating on the DNA strand (polarity was only demonstrated for *S. thermophilus* Cas3). ATP hydrolysis however was differentially stimulated in the two systems, with the *E. coli* Cas3 demonstrating high rates of ATP hydrolysis independent of nucleic acids (in agreement with the general model for DExH-box proteins), while *S. thermophilus* Cas3 required the presence of ssDNA for efficient activity and the rates obtained were much lower than its *E. coli* counterpart. The situation in *S. solfataricus* is even less clear, as the separately encoded HD-nuclease SsoCas3" has demonstrated a substrate specificity for dsDNA and dsRNA but was unable to cleave ssDNA (Han and Krauss, 2009).

Cas3' from *S. solfataricus* was expressed in recombinant form in *E. coli* and purified to homogeneity. Unfortunately, attempts to express SsoCas3" recombinantly were unsuccessful. Co-expression trials of the two proteins in various dual vector systems were also unsuccessful. The eight sequence motifs characteristic of superfamily 2, DExH-box family proteins were detected in the primary sequence of the protein. As expected, SsoCas3' is a multifunctional protein clearly shown to possess ssDNA-dependent ATPase, ATP-independent strand annealing and RNA strand exchange activities. Minimal ATP hydrolysis was detected in the absence of ssDNA, as well as in the presence of ssRNA or any double stranded species including RNA-DNA hybrids. The rate constants obtained were very low in comparison to the *S. thermophilus* Cas3 which was also stimulated by ssDNA. SsoCas3' was able to promote ATP-independent annealing of complementary RNA and DNA strands, as well as the invasion of an RNA strand into a complementary RNA-DNA heteroduplex. Almost 100% annealing of two separate strands in the presence of SsoCas3' was observed in the timeframe of the reaction, as opposed to no annealing in the absence of protein. Regarding the strand exchange activity, in average about 50% of the single RNA strand was converted into heteroduplex product in the presence of SsoCas3' in every assay replicate, in comparison to zero or 5% of strand exchange taking place in

the absence of protein. Invasion of a DNA strand into complementary heteroduplex was not supported by SsoCas3'. Considering the limited results we obtained, and given that additional controls must be carried out in order to characterise fully this activity, on first inspection the data suggest that annealing is probably promoted due to passive binding of nucleic acids by SsoCas3'. Strand rearrangement then is enabled between the proximal strands towards formation of a stable species. This effect was protein dependent, and clearly influenced by increasing temperature, potentially because the RNA-DNA duplexes became more unstable. A first attempt to extend this activity and investigate whether SsoCas3' could promote the formation of *bona fide* R-loops, in the light of the results by Howard *et al.* (2011) yielded encouraging results, and should be investigated further. These results indicated a potential role for the Csa2-Cas5a complex in this activity. This is unsurprising, as this complex is suggested to form the core of the archaeal CASCADE, the bacterial counterpart of which has been shown to promote formation of R-loops in *E. coli* (Jore *et al.* 2011; Howard *et al.* 2011). It can be hypothesised that the promotion of strand rearrangement with a preference for ssRNA could be useful during the scanning of invader sequences for the correct complementary sequence to the aCASCADE-bound crRNA. Local remodeling of RNP complexes is a common activity displayed by DExH/D-box proteins, which could be an additional activity for the helicase domain of Cas3.

The putative ATP-dependent helicase activity of SsoCas3' is ambiguous. As described in the respective section, SsoCas3' could effectively unwind one set of substrates comprising of the CRISPR repeat sequence of locus B, but not substrates with unrelated sequence or a mock physiological substrate consisting of the crRNA and the complementary DNA protospacer. In assays including the first set of substrates, SsoCas3' exhibited a non-processive helicase activity specific for DNA-RNA duplexes, but not for dsDNA or dsRNA substrates. The requirement for an RNA strand is consistent with the characteristics of all DExH/D-box proteins. An apparent lack of directionality resulted in the observation that it could unwind effectively blunt DNA-RNA duplexes, but we were unable to determine whether it translocates on DNA or RNA strands. The helicase activity was ATP-dependent and concentration dependent, suggesting that it is not an artefact. Moreover, inactivation of the Walker A motif by site-directed mutagenesis abolished activity. If we accept that this activity is real, then the observed enzymatic activity of SsoCas3' overall is more reminiscent of the typical activity of DEAD-box proteins, rather than DExH-box proteins (described in 5.1.2). This is where the role of the HD-nuclease must be considered. The indispensable role of additional domains in complementing, regulating and mediating the activities of the basic tandem RecA-like core fold in DExH-box proteins has been discussed previously. A similar point can be made about the role of interacting proteins and co-factors for DEAD-box proteins, which comprise essentially of this basic tandem Rec-A like motor fold. It could be hypothesised that this reflects the

relationship between the HD-nuclease and the helicase domains in Cas3 proteins, either as separate protein entities or as subdomains of the same polypeptide. SsoCas3' comprises mainly of the two motor RecA-like folds, with the additional domain more likely to mediate protein interactions than influence the activity of the basic core, if the tertiary structure prediction by Phyre2 is correct. It is likely therefore that in this state SsoCas3' resembles the topology and domain organisation of DEAD-box proteins, although sequence motif conservation and substrate binding mode certainly differ. This could potentially explain the basic activity demonstrated by the protein at this stage, such as the low ATPase rates, requirement for ssDNA stimulation and the lack of directionality in helicase activity, or the absence of it. Moreover, DEAD-box proteins that act as RNA "chaperones" typically promote a steady-state equilibrium between the two conflicting activities of annealing and unwinding in the presence of ATP, consistent with their role in RNP remodeling (Halls *et al.* 2007). The lack of helicase activity on the other substrate sets is puzzling, as it is unlikely for a helicase of this type to exhibit such strict sequence specificity, especially in the absence of its natural interacting partners. It is possible that the first substrate set of CRISPR repeat sequences is problematic and inherently unstable, and in reality SsoCas3' does not display helicase activity in this form, although many questions are raised in this case as well. Even though we strove to carry out all experiments in triplicate, reliable reaction rates for all activities demonstrated by SsoCas3' could not be obtained as the protein's efficiency and behaviour was different depending on purification batches, potential contaminants, storage conditions and length, and assay components.

Either way, our results until this point indicate that characterisation of SsoCas3' cannot be completed without its key partner, the HD-nuclease SsoCas3". Even though their interdomain interactions are not obvious from the studies of the fused proteins (Sinkunas *et al.* 2011; Howards *et al.* 2011), a physical interaction between the separate proteins is almost certain. On a first level, this interaction would potentially stabilise and alter the conformation of both proteins, with unpredictable effects on their activity. It is possible that this interaction would fine-tune the helicase and annealing activities of SsoCas3', in agreement with the mechanisms of DExH-box helicases. Further research should focus on obtaining SsoCas3" and investigating the activities of both proteins in relation to one another and to the aCASCADE components, in order to be able to elucidate the CRISPR interference mechanism.

Potential protein interactions of recombinant SsoCas3' with other recombinant Cas proteins predicted to be involved in the interference stage (namely the core aCASCADE components and potential subunits) were also investigated and results are presented in Chapter 4. No stable protein interactions were observed for SsoCas3', although the possibility of transient interactions cannot be ruled out. It is

thought that the CASCADE complex somehow recruits Cas3 to catalyse final cleavage of invader DNA, although the details of this mechanism are yet to be determined.

Chapter 6

Conclusions and future work

This thesis described the initial purification and characterisation of some of the key elements in CRISPR-mediated antiviral defence in *S. solfataricus*. The complexity of the CRISPR/Cas system in this organism, containing two different subtypes (I-A and III-B) with almost 60 *cas* genes organised in multiple operons is both a challenge and an advantage. The existence of multiple reasonably conserved orthologues for a given *cas* gene means that multiple candidates are available for solubility screening and crystallisation trials, a practical approach which proved extremely valuable during the course of this study. On the other hand, this situation makes genetic studies extremely complicated to interpret at a mechanistic level, and knockout studies almost impossible.

A general observation made during this study was that most of the Cas proteins in *S. solfataricus* were extremely unstable during heterologous expression in *E. coli* or not expressed at all. This is not unexpected since the completion of this pathway requires a number of tightly regulated protein-protein interactions, and most of these proteins are predicted to or have been shown in other organisms to form small or large protein complexes (e.g. Hale *et al.* 2009; Brouns *et al.* 2008). Individual protein expression in this case fails due to the lack of the appropriate partner, which may stabilise the protein and in many cases alter its activity. Further effort needs to be put on towards overcoming the expression and solubility problems with the Cas proteins of *S. solfataricus*, and apart from co-expression studies which were attempted and in some cases successful (for example for the Csa2-Cas5a complex), it seems that the most promising strategy is the over-expression of recombinant proteins in their native strain, by using the genetic systems developed for *S. solfataricus* by S. V. Albers (Albers *et al.* 2006). This would enable protein production in their native environment, where they can acquire any post-translational modifications potentially necessary for protein function, and where correct protein folding and protein stability is ensured by the endogenous expression of their appropriate partners. This strategy has been shown to improve the yield and enzymatic activity of over-expressed proteins in a number of cases, and can also be used to identify novel interacting partners by pull-down experiments, in a method similar to the one

employed here for the native aCASCADE and the Cmr complex. This work is ongoing and is being carried out by Dr Christophe Rouillon and Dr Jing Zhang (Professor Malcolm White's group, St Andrews University).

The attempt to reconstruct the *S. solfataricus* Cmr complex is a case in point, as only four (Cmr1, Cmr3, Cmr4 and Cmr7) out of the six subunits were successfully expressed in recombinant form. Apart from the pairwise interaction between Cmr1 and Cmr3 and the nucleic acid binding ability of Cmr1, little information was gained by the biochemical study of these proteins individually. In the absence of the predicted catalytic subunits of the Cmr complex (Cas10/Cmr2 and Cmr5, Makarova *et al.* 2011b), it is evident that a meaningful characterisation of their role within the complex or of the complex itself cannot be achieved.

The antibody-assisted isolation of the native Cmr complex from *S. solfataricus* described in Chapter 3 confirmed the constituent production of this complex under normal (not virus infected) growth conditions, and is a first step towards understanding its characteristics and function. We were unable to detect a nuclease or polymerase activity with this native complex, but the elucidation of the homologous complex in *P. furiosus* indicated that our initial hypothesis was mistaken and further investigation is needed. The optimisation of the isolation process or the native overexpression conditions in order to increase the final yield is crucial for further experimentation. This will also enable structural studies of the Cmr complex, either by TEM, SAXS analysis or classic crystallography, which will be valuable in understanding the target recognition and interference mechanism. Native mass spectrometry should also be attempted in order to determine the exact stoichiometry of the complex, as demonstrated for the *E. coli* CASCADE and the *P. aeruginosa* Csy complex (Jore *et al.* 2011; Wiedenheft *et al.* 2011). A small progress in this matter was made with the elucidation of the structure of Cmr7, described in chapter 3 and potentially providing the structural scaffold of the Cmr complex, as indicates its overrepresentation in the purified sample.

It will be interesting to determine whether the Cmr complex in *S. solfataricus* has the same functionality as the *P. furiosus* complex, namely that it mediates target RNA cleavage guided by the bound crRNA via a molecular ruler mechanism *in vitro* (Hale *et al.* 2009). The catalytic subunit in *P. furiosus* was not identified, but it is predicted to be either the HD domain of Cas10, or Cmr5 (Makarova *et al.* 2011b). Site-directed mutagenesis could be employed to determine this in the case of a similar activity of the SsoCmr, provided there is a method to overexpress the mutant proteins. Whether the targeting of RNA is also taking place *in vivo* is still unclear. Interestingly, the type III-A system of *S. epidermidis* has been shown to target DNA *in vivo* (Marraffini and Sontheimer, 2008) and shares a similar Cas protein composition with

type III-B systems. Moreover, no archaeal RNA viruses are currently known, making it difficult to speculate about the biological relevance of this function.

Chapter 4 described the purification and characterisation of a stable recombinant complex composed of Csa2 and Cas5a, two core proteins of type I-A systems. These proteins belong to families Cas7 and Cas5 and are orthologous to the CasC and CasD subunits of the *E. coli* CASCADE. It is demonstrated that Csa2 and Cas5a form the stable core of an archaeal CASCADE-like complex for antiviral defence, termed aCASCADE. The recombinant Csa2-Cas5a complex is able to specifically bind crRNA and recognise complementary target ssDNA *in vitro*. Further studies are needed to fully characterise the recognition mechanism and substrate requirements, for example whether it can recognise and bind a complementary target in a dsDNA substrate, in a mechanism similar to the *E. coli* CASCADE. Mass spectrometry analysis of the native aCASCADE expressed in *S. solfataricus* by N. Lintner in the group of Martin Lawrence in Montana State University Bozeman demonstrated that the aCASCADE consists of accessory subunits apart from the core Csa2-Cas5a complex, namely Csa5, Cas6 and perhaps Csa4 (Cas8a2), but these proteins are either not forming a stable interaction with the core complex or are present in sub-stoichiometric amounts. Nucleic acid extraction from the native complex carried out by N. Lintner revealed that aCASCADE co-purifies with the mature form of crRNA, which consists of the conserved elements identified in other organisms (Brouns *et al.* 2008), namely the repeat derived 8 nt 5' handle and 21 nt 3' handle interspersed by a spacer sequence. This was reminiscent of the products of pre-CRISPR transcript processing by Cas6 in *P. furiosus*, an activity which was verified here for the *S. solfataricus* Cas6.

Transmission electron microscopy was employed to identify the structural characteristics of the native aCASCADE by N. Lintner. The formation of extended right-handed helices was observed, leading to the proposal of a structural model for the assembly of aCASCADE by N. Lintner and M. Lawrence. According to this model, aCASCADE under physiological conditions (not Csa2 overexpression) would adopt a semi-helical arch-shaped structure with a backbone composed by multiple (6-8) Csa2 subunits and a single crRNA bound along the length of the arch. Cas5a and the accessory subunits could potentially nucleate and regulate the growth of the assembly, apart from mediating specific nucleic acid or protein interactions (Lintner *et al.* 2011). Taking into account the emerging structures for CASCADE and the Csy complex (Jore *et al.* 2011; Wiedenheft *et al.* 2011), we can observe the conservation of an arch-shaped core structure in all complexes, indicating its importance and presumably that it represents the core structure of all type I CASCADE-like complexes.

Cas3 has been shown to be required for the final target cleavage and invader silencing in type I-E systems such as *E. coli* (Brouns *et al.* 2011). In type I-A systems

the identified SF2 helicase and HD nuclease motifs of this protein are encoded as two separate proteins, Cas3' and Cas3". The purification and initial characterisation of Cas3' from *S. solfataricus* is described in chapter 5. The existence of DExH-box family motifs is generally associated with ATP-dependent RNP remodelling (Pyle, 2008), and does not necessarily imply a helicase activity. Therefore, the helicase, ATPase and strand exchange activities of SsoCas3' which contains all the canonical DExH-box family signature motifs were investigated biochemically in this chapter. SsoCas3' displayed a weak ssDNA-stimulated ATPase activity and an ability to promote strand exchange between single strand RNA and hybrid RNA-DNA duplexes. The results obtained from investigation of the helicase activity of SsoCas3' were contradictory. With one set of substrates consisting of CRISPR repeat sequences SsoCas3' exhibited the ability to unwind blunt DNA-RNA hybrids, but not dsDNA or dsRNA, in an ATP-dependent matter. However, we were unable to confirm this activity with different substrates, and therefore would be premature to characterise the protein as a *de facto* helicase. Moreover, it is not understood why the ATPase activity of SsoCas3' was not stimulated by RNA-DNA hybrid duplexes but by ssDNA, unless the function of this protein involves initial recognition of a single strand DNA region in order to carry out remodelling of an RNA-DNA hybrid, such as the one formed upon recognition of the DNA protospacer target by the crRNA-aCASCADE complex. Initial studies into the possibility of R-loop formation by SsoCas3' and the Csa2-Cas5a complex gave positive results, indicating that a mechanism similar to the R-loop formation observed for *E. coli* CASCADE (Jore *et al.* 2011) and the *E. coli* Cas3 (Howard *et al.* 2011) could perhaps be taking place in *S. solfataricus* as well. Further experimentation is needed to clarify this matter.

The situation of Cas3' and Cas3" is another case that would benefit greatly from native expression in *S. solfataricus*. As discussed in Chapter 5, the behaviour of characterised Cas3 orthologues differs greatly from the behaviour exhibited here by SsoCas3' and the behaviour of the individual HD-domain protein Cas3" from *S. solfataricus* (Han *et al.* 2009). Since it is predicted almost certainly that these two proteins interact, and considering the role that accessory domains and interacting partners play in regulating the function of DExH/D-box helicases (Pyle, 2008), it becomes apparent that a physiologically relevant functional characterisation of any of these proteins cannot proceed without the other.

Expression of SsoCas3" is also essential to verify its predicted role along with Cas3' in the final step of CRISPR/Cas interference in cooperation with the aCASCADE. Genetic studies in *E. coli* and biochemical characterisation of the type I-E enzyme (Brouns *et al.* 2008; Sinkunas *et al.* 2011; Howard *et al.* 2011) have led to the proposal that Cas3 is recruited by CASCADE to the site of crRNA-mediated protospacer recognition. Once present, it can perform the final cleavage of both strands of the

dsDNA target and perhaps catalyse localised remodeling of the CASCADE-crRNA-DNA complex in order to release the crRNA and/or release the DNA strand for cleavage. Interaction between SsoCas3' and the Csa2-Cas5a complex or the native aCASCADE in *S. solfataricus* was not detected during this study, but considering the current unavailability of all proteins predicted to be involved that is hardly surprising. An explanation could also be that that an interaction might only take place under certain conditions or other signaling molecules might also be involved.

A feature of both native complexes characterised in this study (Cmr complex and aCASCADE) was their constitutive expression, indicated by the fact that it was possible to isolate them from crude cell lysate under normal growth conditions (for aCASCADE, see Lintner *et al.* 2011). This is a situation observed in all archaeal systems studied to date (Tang *et al.* 2005; Hale *et al.* 2009; Lillestol *et al.* 2009), suggesting that in these systems CRISPR/Cas is not under negative regulation as opposed to the situation observed in bacteria. This is in agreement with the system being in “surveillance” mode, constantly alert for potential extrachromosomal infections. The co-existence of types I-A and III-B in *S. solfataricus* indicates that this organism can target both DNA and RNA, widening the range of immunity. The prevalence of type III systems in thermophilic archaea perhaps reflects an adaptation to a specific need of these organisms, namely the need to silence some form of RNA. The source of this target RNA is yet to be determined, but it can be certain that the polymorphism and diversity of the archaeal virosphere will continue to surprise us.

References

- Agari Y, Sakamoto K, Tamakoshi M, Oshima T, Kuramitsu S, Shinkai A (2010) Transcription profile of *Thermus thermophilus* CRISPR systems after phage infection. *J Mol Biol* **395**: 270-281
- Agari Y, Yokoyama S, Kuramitsu S, Shinkai A (2008) X-ray crystal structure of a CRISPR-associated protein, Cse2, from *Thermus thermophilus* HB8. *Proteins* **73**: 1063-1067
- Agrawal N, Dasaradhi PVN, Mohammed A, Malhotra P, Bhatnagar RK, Mukherjee SK (2003) RNA interference: biology, mechanism, and applications. *Microbiol Mol Biol Rev* **67**: 657-685
- Al-Attar S, Westra ER, van der Oost J, Brouns SJJ (2011) Clustered regularly interspaced short palindromic repeats (CRISPRs): the hallmark of an ingenious antiviral defence mechanism in prokaryotes. *Biol Chem* **392**: 277-289
- Albers SV, Jonuscheit M, Dinkelaker S, Urich T, Kletzin A, Tampe R, Driessen AJ, Schleper C (2006) Production of recombinant and tagged proteins in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Applied and Environmental Microbiology* **72**: 102-111
- Aliyari R, Ding S-W (2009) RNA-based viral immunity initiated by the Dicer family of host immune receptors. *Immunol Rev* **227**: 176-188
- Anantharaman V, Iyer LM, Aravind L (2010) Presence of a classical RRM-fold palm domain in Thg1-type 3'-5' nucleic acid polymerases and the origin of the GGDEF and CRISPR polymerase domains. *Biol Direct* **5**: 43
- Andersen CBF, Ballut L, Johansen JS, Chamieh H, Nielsen KH, Oliveira CLP, Pedersen JS, Séraphin B, Le Hir H, Andersen GR (2006) Structure of the exon junction core complex with a trapped DEAD-box ATPase bound to RNA. *Science* **313**: 1968-1972
- Anderson RE, Brazelton WJ, Baross JA (2011) Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. *FEMS Microbiology Ecology* **77**: 120-133
- Andersson AF, Banfield JF (2008) Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* **320**: 1047-1050
- Aravind L, Koonin EV (1998) The HD domain defines a new superfamily of metal-dependent phosphohydrolases. *Trends Biochem Sci* **23**: 469-472

- Arcus VL, Mckenzie JL, Robson J, Cook GM (2011) The PIN-domain ribonucleases and the prokaryotic VapBC toxin-antitoxin array. *Protein Engineering Design and Selection* **24**: 33-40
- Babu M, Beloglazova N, Flick R, Graham C, Skarina T, Nocek B, Gagarinova A, Pogoutse O, Brown G, Binkowski A, Phanse S, Joachimiak A, Koonin EV, Savchenko A, Emili A, Greenblatt J, Edwards AM, Yakunin AF (2011) A dual function of the CRISPR-Cas system in bacterial antiviral immunity and DNA repair. *Mol Microbiol* **79**: 484-502
- Bagasra O, Prilliman KR (2004) RNA interference: the molecular immune system. *J Mol Histol* **35**: 545-553
- Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* **34**: W369-373
- Baker NA (2004) Poisson-Boltzmann methods for biomolecular electrostatics. *Methods Enzymol* **383**: 94-118
- Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 10037-10041
- Banroques J, Doère M, Dreyfus M, Linder P, Tanner NK (2010) Motif III in superfamily 2 "helicases" helps convert the binding energy of ATP into a high-affinity RNA binding site in the yeast DEAD-box protein Ded1. *J Mol Biol* **396**: 949-966
- 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
- Beloglazova N, Brown G, Zimmerman MD, Proudfoot M, Makarova KS, Kudritska M, Kochinyan S, Wang S, Chruszcz M, Minor W, Koonin EV, Edwards AM, Savchenko A, Yakunin AF (2008) A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *The Journal of Biological Chemistry* **283**: 20361-20371
- Bizebard T, Ferlenghi I, Iost I, Dreyfus M (2004) Studies on Three E. coli DEAD-Box Helicases Point to an Unwinding Mechanism Different from that of Model DNA Helicases †. *Biochemistry* **43**: 7857-7866
- Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P (2007) CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**: 209
- Blower TR, Fineran PC, Johnson MJ, Toth IK, Humphreys DP, Salmond GPC (2009) Mutagenesis and functional characterization of the RNA and protein components of the toxIN abortive infection and toxin-antitoxin locus of *Erwinia*. *Journal of bacteriology* **191**: 6029-6039
- Blower TR, Pei XY, Short FL, Fineran PC, Humphreys DP, Luisi BF, Salmond GPC (2011) A processed noncoding RNA regulates an altruistic bacterial antiviral system. *Nature Structural & Molecular Biology* **18**: 185
- Blower TR, Salmond GP, Luisi BF (2011) Balancing at survival's edge: the structure and adaptive benefits of prokaryotic toxin - antitoxin partners. *Curr Opin Struct Biol* **21**: 109-118

- 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
- Bowers HA (2006) Discriminatory RNP remodeling by the DEAD-box protein DED1. *RNA* **12**: 903-912
- Brochier-Armanet C, Forterre P, Gribaldo S (2011) Phylogeny and evolution of the Archaea: one hundred genomes later. *Current Opinion in Microbiology* **14**: 274-281
- Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJH, Snijders APL, Dickman MJ, Makarova KS, Koonin EV, van der Oost J (2008) Small CRISPR RNAs guide antiviral defence in prokaryotes. *Science* **321**: 960-964
- Browne EP, Li J, Chong M, Littman DR (2005) Virus-host interactions: new insights from the small RNA world. *Genome Biol* **6**: 238
- Buchon N, Vaury C (2006) RNAi: a defensive RNA-silencing against viruses and transposable elements. *Heredity* **96**: 195-202
- Bukowski M, Rojowska A, Wladyka B (2011) Prokaryotic toxin-antitoxin systems - the role in bacterial physiology and application in molecular biology. *Acta Biochim Pol* **58**: 1-9
- Cady KC, O'Toole GA (2011) Non-identity Targeting of Yersinia-Subtype CRISPR-Prophage Interaction Requires the Csy and Cas3 Proteins. *Journal of Bacteriology* **193**: 3433-3445
- Cady KC, White AS, Hammond JH, Abendroth MD, Karthikeyan RSG, Lalitha P, Zegans ME, O'Toole GA (2011) Prevalence, conservation and functional analysis of Yersinia and Escherichia CRISPR regions in clinical Pseudomonas aeruginosa isolates. *Microbiology* **157**: 430-437
- Campo MD, Mohr S, Jiang Y, Jia H, Jankowsky E, Lambowitz AM (2009) Unwinding by Local Strand Separation Is Critical for the Function of DEAD-Box Proteins as RNA Chaperones. *J Mol Biol* **389**: 674-693
- Carte J, Pfister NT, Compton MM, Terns RM, Terns MP (2010) Binding and cleavage of CRISPR RNA by Cas6. *RNA* **16**: 2181-2188
- Carte J, Wang R, Li H, Terns RM, Terns MP (2008) Cas6 is an endoribonuclease that generates guide RNAs for invader defence in prokaryotes. *Genes Dev* **22**: 3489-3496
- Carthew RW, Sontheimer EJ (2009) Origins and Mechanisms of miRNAs and siRNAs. *Cell* **136**: 642-655
- Caruthers JM, Johnson ER, McKay DB (2000) Crystal structure of yeast initiation factor 4A, a DEAD-box RNA helicase. *Proceedings of the National Academy of Sciences of the United States of America* **97**: 13080-13085
- Caruthers JM, McKay DB (2002) Helicase structure and mechanism. *Curr Opin Struct Biol* **12**: 123-133
- Chakraborty S, Waise TMZ, Hassan F, Kabir Y, Smith MA, Arif M (2009) Assessment of the evolutionary origin and possibility of CRISPR-Cas (CASS) mediated RNA interference pathway in Vibrio cholerae O395. *In Silico Biol (Gedruckt)* **9**: 245-254

- Chopin M-C, Chopin A, Bidnenko E (2005) Phage abortive infection in lactococci: variations on a theme. *Curr Opin Microbiol* **8**: 473-479
- Collins RE, Cheng X (2006) Structural and biochemical advances in mammalian RNAi. *J Cell Biochem* **99**: 1251-1266
- Comeau AM, Krisch HM (2005) War is peace--dispatches from the bacterial and phage killing fields. *Curr Opin Microbiol* **8**: 488-494
- Cooper Charlotte R, Daugherty Amanda J, Tachdjian S, Blum Paul H, Kelly Robert M (2009) Role of vapBC toxin-antitoxin loci in the thermal stress response of *Sulfolobus solfataricus*. *Biochem Soc Trans* **37**: 123
- Cordin O, Banroques J, Tanner NK, Linder P (2006) The DEAD-box protein family of RNA helicases. *Gene* **367**: 17-37
- Cui Y, Li Y, Gorgé O, Platonov ME, Yan Y, Guo Z, Pourcel C, Dentovskaya SV, Balakhonov SV, Wang X, Song Y, Anisimov AP, Vergnaud G, Yang R (2008) Insight into microevolution of *Yersinia pestis* by clustered regularly interspaced short palindromic repeats. *PLoS ONE* **3**: e2652
- Dagan T, Artzy-Randrup Y, Martin W (2008) Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proceedings of the National Academy of Sciences of the United States of America* **105**: 10039-10044
- De Felice M, Aria V, Esposito L, De Falco M, Pucci B, Rossi M, Pisani FM (2007) A novel DNA helicase with strand-annealing activity from the crenarchaeon *Sulfolobus solfataricus*. *The Biochemical journal* **408**: 87-95
- de la Cruz J, Kressler D, Linder P (1999) Unwinding RNA in *Saccharomyces cerevisiae*: DEAD-box proteins and related families. *Trends Biochem Sci* **24**: 192-198
- DeBoy RT, Mongodin EF, Emerson JB, Nelson KE (2006) Chromosome evolution in the Thermotogales: large-scale inversions and strain diversification of CRISPR sequences. *Journal of Bacteriology* **188**: 2364-2374
- Del Campo M, Tijerina P, Bhaskaran H, Mohr S, Yang Q, Jankowsky E, Russell R, Lambowitz AM (2007) Do DEAD-box proteins promote group II intron splicing without unwinding RNA? *Mol Cell* **28**: 159-166
- Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, Charpentier E (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**: 602-607
- Deveau H, Barrangou R, Garneau JE, Labonté J, Fremaux C, Boyaval P, Romero DA, Horvath P, Moineau S (2008) Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *Journal of Bacteriology* **190**: 1390-1400
- Deveau H, Garneau JE, Moineau S (2010) CRISPR/Cas system and its role in phage-bacteria interactions. *Annu Rev Microbiol* **64**: 475-493
- Díez-Villaseñor C, Almendros C, García-Martínez J, Mojica FJM (2010) Diversity of CRISPR loci in *Escherichia coli*. *Microbiology* **156**: 1351-1361

- Dumont S, Cheng W, Serebrov V, Beran RK, Tinoco I, Pyle AM, Bustamante C (2006) RNA translocation and unwinding mechanism of HCV NS3 helicase and its coordination by ATP. *Nature* **439**: 105-108
- Dyall-Smith M (2011) Dangerous weapons: a cautionary tale of CRISPR defence. *Mol Microbiol* **79**: 3-6
- Ebihara A, Yao M, Masui R, Tanaka I, Yokoyama S, Kuramitsu S (2006) Crystal structure of hypothetical protein TTHB192 from *Thermus thermophilus* HB8 reveals a new protein family with an RNA recognition motif-like domain. *Protein Sci* **15**: 1494-1499
- Edgar R, Qimron U (2010) The *Escherichia coli* CRISPR system protects from λ lysogenization, lysogens, and prophage induction. *Journal of Bacteriology* **192**: 6291-6294
- Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113
- Edgar RC (2007) PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* **8**: 18
- Edgar RC, Batzoglou S (2006) Multiple sequence alignment. *Curr Opin Struct Biol* **16**: 368-373
- Edgar RC, Sjölander K (2004) COACH: profile-profile alignment of protein families using hidden Markov models. *Bioinformatics* **20**: 1309-1318
- Fairman ME (2004) Protein Displacement by DExH/D "RNA Helicases" Without Duplex Unwinding. *Science* **304**: 730-734
- Fairman-Williams ME, Guenther U-P, Jankowsky E (2010) SF1 and SF2 helicases: family matters. *Curr Opin Struct Biol* **20**: 313-324
- Fineran PC, Blower TR, Foulds IJ, Humphreys DP, Lilley KS, Salmond GPC (2009) The phage abortive infection system, ToxIN, functions as a protein-RNA toxin-antitoxin pair. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 894-899
- Fire A (1999) RNA-triggered gene silencing. *Trends Genet* **15**: 358-363
- Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**: 806-811
- Garneau JE, Dupuis M-È, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadán AH, Moineau S (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**: 67-71
- Garrett RA, Prangishvili D, Shah SA, Reuter M, Stetter KO, Peng X (2010) Metagenomic analyses of novel viruses and plasmids from a cultured environmental sample of hyperthermophilic neutrophiles. *Environ Microbiol* **12**: 2918-2930
- Garrett RA, Shah SA, Vestergaard G, Deng L, Gudbergdottir S, Kenchappa CS, Erdmann S, She Q (2011) CRISPR-based immune systems of the *Sulfolobales*: complexity and diversity. *Biochem Soc Trans* **39**: 51-57

- Gesner EM, Schellenberg MJ, Garside EL, George MM, Macmillan AM (2011) Recognition and maturation of effector RNAs in a CRISPR interference pathway. *Nature Structural & Molecular Biology* **18**: 688-692
- Gill EE, Brinkman FSL (2011) The proportional lack of archaeal pathogens: Do viruses/phages hold the key? *Bioessays* **33**: 248-254
- Godde JS, Bickerton A (2006) The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J Mol Evol* **62**: 718-729
- Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* **19**: 2226-2238
- Grissa I, Bouchon P, Pourcel C, Vergnaud G (2008) On-line resources for bacterial micro-evolution studies using MLVA or CRISPR typing. *Biochimie* **90**: 660-668
- 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
- Grissa I, Vergnaud G, Pourcel C (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* **35**: W52-57
- Grissa I, Vergnaud G, Pourcel C (2008) CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* **36**: W145-148
- Grohman JK, Del Campo M, Bhaskaran H, Tijerina P, Lambowitz AM, Russell R (2007) Probing the mechanisms of DEAD-box proteins as general RNA chaperones: the C-terminal domain of CYT-19 mediates general recognition of RNA. *Biochemistry* **46**: 3013-3022
- Gudbergdottir S, Deng L, Chen Z, Jensen JVK, Jensen LR, She Q, Garrett RA (2011) Dynamic properties of the *Sulfolobus* CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. *Mol Microbiol* **79**: 35-49
- Guinane CM, Kent RM, Norberg S, Hill C, Fitzgerald GF, Stanton C, Ross RP (2011) Host Specific Diversity in *Lactobacillus johnsonii* as Evidenced by a Major Chromosomal Inversion and Phage Resistance Mechanisms. *PLoS ONE* **6**: e18740
- Guo L, Brügger K, Liu C, Shah SA, Zheng H, Zhu Y, Wang S, Lillestøl RK, Chen L, Frank J, Prangishvili D, Paulin L, She Q, Huang L, Garrett RA (2011) Genome analyses of icelandic strains of *Sulfolobus islandicus*, model organisms for genetic and virus-host interaction studies. *Journal of Bacteriology* **193**: 1672-1680
- Haft DH, Selengut J, Mongodin EF, Nelson KE (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, Kleppe K, Terns RM, Terns MP (2008) Prokaryotic silencing (psi)RNAs in *Pyrococcus furiosus*. *RNA* **14**: 2572-2579

- 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
- Hall MC, Matson SW (1999) Helicase motifs: the engine that powers DNA unwinding. *Mol Microbiol* **34**: 867-877
- Halls C, Mohr S, Del Campo M, Yang Q, Jankowsky E, Lambowitz AM (2007) Involvement of DEAD-box proteins in group I and group II intron splicing. Biochemical characterization of Mss116p, ATP hydrolysis-dependent and -independent mechanisms, and general RNA chaperone activity. *J Mol Biol* **365**: 835-855
- Hammond SM (2005) Dicing and slicing: the core machinery of the RNA interference pathway. *FEBS Lett* **579**: 5822-5829
- Han D, Krauss G (2009) Characterization of the endonuclease SSO2001 from *Sulfolobus solfataricus* P2. *FEBS Lett* **583**: 771-776
- Han D, Lehmann K, Krauss G (2009) SSO1450--a CAS1 protein from *Sulfolobus solfataricus* P2 with high affinity for RNA and DNA. *FEBS Lett* **583**: 1928-1932
- Hannon GJ (2002) RNA interference. *Nature* **418**: 244-251
- Hardin JW, Hu YX, McKay DB (2010) Structure of the RNA Binding Domain of a DEAD-Box Helicase Bound to Its Ribosomal RNA Target Reveals a Novel Mode of Recognition by an RNA Recognition Motif. *J Mol Biol* **402**: 412-427
- Haurwitz RE, Jinek M, Wiedenheft B, Zhou K, Doudna JA (2010) Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* **329**: 1355-1358
- Hazan R, Engelberg-Kulka H (2004) *Escherichia coli* mazEF-mediated cell death as a defence mechanism that inhibits the spread of phage P1. *Mol Genet Genomics* **272**: 227-234
- He J, Deem MW (2010) Heterogeneous diversity of spacers within CRISPR (clustered regularly interspaced short palindromic repeats). *Phys Rev Lett* **105**: 128102
- Heidelberg JF, Nelson WC, Schoenfeld T, Bhaya D (2009) Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. *PLoS ONE* **4**: e4169
- Held NL, Herrera A, Cadillo-Quiroz H, Whitaker RJ (2010) CRISPR associated diversity within a population of *Sulfolobus islandicus*. *PLoS ONE* **5**
- Held NL, Whitaker RJ (2009) Viral biogeography revealed by signatures in *Sulfolobus islandicus* genomes. *Environ Microbiol* **11**: 457-466
- Holm L, Rosenstrom P (2010) Dali server: conservation mapping in 3D. *Nucleic Acids Research* **38**: W545-549
- Hopfner K-P, Michaelis J (2007) Mechanisms of nucleic acid translocases: lessons from structural biology and single-molecule biophysics. *Curr Opin Struct Biol* **17**: 87-95

- Hornung V, Ellegast J, Kim S, Brzozka K, Jung A, Kato H, Poeck H, Akira S, Conzelmann K-K, Schlee M, Endres S, Hartmann G (2006) 5'-Triphosphate RNA Is the Ligand for RIG-I. *Science* **314**: 994-997
- Horvath P, Barrangou R (2010) CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**: 167-170
- Horvath P, Coûté-Monvoisin A-C, Romero DA, Boyaval P, Fremaux C, Barrangou R (2009) Comparative analysis of CRISPR loci in lactic acid bacteria genomes. *Int J Food Microbiol* **131**: 62-70
- Horvath P, Romero DA, Coûté-Monvoisin A-C, Richards M, Deveau H, Moineau S, Boyaval P, Fremaux C, Barrangou R (2008) Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *Journal of Bacteriology* **190**: 1401-1412
- Hoskisson PA, Smith MCM (2007) Hypervariation and phase variation in the bacteriophage 'resistome'. *Curr Opin Microbiol* **10**: 396-400
- Howard JL, Delmas S, Ivancic-Bace I, Bolt EL (2011) Helicase dissociation and annealing of RNA-DNA hybrids by *Escherichia coli* Cas3 protein. *The Biochemical journal* (in process)
- Huang Y (2002) The ATPase, RNA Unwinding, and RNA Binding Activities of Recombinant p68 RNA Helicase. *Journal of Biological Chemistry* **277**: 12810-12815
- Hutvagner G, Simard MJ (2008) Argonaute proteins: key players in RNA silencing. *Nat Rev Mol Cell Biol* **9**: 22-32
- Hyman P, Abedon ST (2010) Bacteriophage host range and bacterial resistance. *Adv Appl Microbiol* **70**: 217-248
- Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A (1987) Nucleotide sequence of the *iap* gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *Journal of Bacteriology* **169**: 5429-5433
- Jankowsky E (2001) Active Disruption of an RNA-Protein Interaction by a DExH/D RNA Helicase. *Science* **291**: 121-125
- Jankowsky E (2005) Biophysics: helicase snaps back. *Nature* **437**: 1245
- Jankowsky E (2006) Remodeling of ribonucleoprotein complexes with DExH/D RNA helicases. *Nucleic Acids Res* **34**: 4181-4188
- Jankowsky E (2007) Biochemistry: indifferent chaperones. *Nature* **449**: 999-1000
- Jankowsky E, Fairman ME (2007) RNA helicases--one fold for many functions. *Curr Opin Struct Biol* **17**: 316-324
- Jankowsky E, Gross CH, Shuman S, Pyle AM (2000) The DExH protein NPH-II is a processive and directional motor for unwinding RNA. *Nature* **403**: 447-451
- Jankowsky E, Jankowsky A (2000) The DExH/D protein family database. *Nucleic Acids Res* **28**: 333-334

- Jansen R, Embden JDAv, Gaastra W, Schouls LM (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **43**: 1565-1575
- Jarmoskaite I, Russell R (2011) DEAD-box proteins as RNA helicases and chaperones. *WIREs RNA* **2**: 135-152
- Jaronczyk K, Carmichael JB, Hobman TC (2005) Exploring the functions of RNA interference pathway proteins: some functions are more RISCy than others? *Biochem J* **387**: 561-571
- Jore MM, Lundgren M, van Duijn E, Bultema JB, Westra ER, Waghmare SP, Wiedenheft B, Pul U, Wurm R, Wagner R, Beijer MR, Barendregt A, Zhou K, Snijders APL, Dickman MJ, Doudna JA, Boekema EJ, Heck AJR, van der Oost J, Brouns SJJ (2011) Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nature Structural & Molecular Biology* **18**: 529-536
- Karginov FV, Hannon GJ (2010) The CRISPR system: small RNA-guided defence in bacteria and archaea. *Mol Cell* **37**: 7-19
- Kato H, Takeuchi O, Sato S, Yoneyama M, Yamamoto M, Matsui K, Uematsu S, Jung A, Kawai T, Ishii KJ, Yamaguchi O, Otsu K, Tsujimura T, Koh C-S, Reis E Sousa C, Matsuura Y, Fujita T, Akira S (2006) Differential roles of MDA5 and RIG-I helicases in the recognition of RNA viruses. *Nature* **441**: 101-105
- Kawaoka J (2005) Choosing between DNA and RNA: the polymer specificity of RNA helicase NPH-II. *Nucleic Acids Res* **33**: 644-649
- Kawaoka J, Jankowsky E, Pyle AM (2004) Backbone tracking by the SF2 helicase NPH-II. *Nature Structural & Molecular Biology* **v11**: 526-530
- Kelley LA, Sternberg MJ (2009) Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc* **4**: 363-371
- Kobayashi I (2001) Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res* **29**: 3742-3756
- Koonin EV, Makarova KS (2009) CRISPR-Cas: an adaptive immunity system in prokaryotes. *F1000 Biol Rep* **1**: 95
- Koonin EV, Wolf YI (2009) Is evolution Darwinian or/and Lamarckian? *Biol Direct* **4**: 42
- Kunin V, Sorek R, Hugenholtz P (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol* **8**: R61
- Labrie SJ, Samson JE, Moineau S (2010) Bacteriophage resistance mechanisms. *Nat Rev Microbiol* **8**: 317-327
- Lawrence CM, White MF (2011) Recognition of archaeal CRISPR RNA: No P in the alindromic repeat? *Structure* **19**: 142-144
- Lecellier C-H, Dunoyer P, Arar K, Lehmann-Che J, Eyquem S, Himber C, Saïb A, Voinnet O (2005) A cellular microRNA mediates antiviral defence in human cells. *Science* **308**: 557-560
- Levin BR (2010) Nasty viruses, costly plasmids, population dynamics, and the conditions for establishing and maintaining CRISPR-mediated adaptive immunity in bacteria. *PLoS Genet* **6**: e1001171

- Levin MK, Gurjar M, Patel SS (2005) A Brownian motor mechanism of translocation and strand separation by hepatitis C virus helicase. *Nature Structural & Molecular Biology* **12**: 429-435
- Lillestøl RK, Redder P, Garrett RA, Brügger K (2006) A putative viral defence mechanism in archaeal cells. *Archaea* **2**: 59-72
- Lillestøl RK, Shah SA, Brügger K, Redder P, Phan H, Christiansen J, Garrett RA (2009) CRISPR families of the crenarchaeal genus *Sulfolobus*: bidirectional transcription and dynamic properties. *Mol Microbiol* **72**: 259-272
- Linder P, Lasko P (2006) Bent out of shape: RNA unwinding by the DEAD-box helicase Vasa. *Cell* **125**: 219-221
- Lingel A, Sattler M (2005) Novel modes of protein-RNA recognition in the RNAi pathway. *Curr Opin Struct Biol* **15**: 107-115
- Lintner NG, Frankel KA, Tsutakawa SE, Alsbury DL, Copié V, Young MJ, Tainer JA, Lawrence CM (2011) The structure of the CRISPR-associated protein Csa3 provides insight into the regulation of the CRISPR/Cas system. *J Mol Biol* **405**: 939-955
- Lintner NG, Kerou M, Brumfield SK, Graham S, Liu H, Naismith JH, Sdano M, Peng N, She Q, Copie V, Young MJ, White MF, Lawrence CM (2011) Structural and functional characterization of an archaeal CASCADE complex for CRISPR-mediated viral defence. *The Journal of Biological Chemistry* **286**:21643-56
- Liu F, Putnam A, Jankowsky E (2008) ATP hydrolysis is required for DEAD-box protein recycling but not for duplex unwinding. *Proceedings of the National Academy of Sciences of the United States of America* **105**: 20209-20214
- Liu H, Rudolf J, Johnson KA, McMahon SA, Oke M, Carter L, McRobbie A-M, Brown SE, Naismith JH, White MF (2008) Structure of the DNA repair helicase XPD. *Cell* **133**: 801-812
- Liu H, Naismith JH (2009) A simple and efficient expression and purification system using two newly constructed vectors. *Protein Expr Purif* **63**:102-11
- Lorsch JR, Herschlag D (1998) The DEAD Box Protein eIF4A. 1. A Minimal Kinetic and Thermodynamic Framework Reveals Coupled Binding of RNA and Nucleotide †. *Biochemistry* **37**: 2180-2193
- Lüking A, Stahl U, Schmidt U (1998) The Protein Family of RNA Helicases. *Critical Reviews in Biochemistry and Molecular Biology* **33**: 259-296
- Machwe A, Xiao L, Groden J, Matson SW, Orren DK (2005) RecQ family members combine strand pairing and unwinding activities to catalyze strand exchange. *The Journal of Biological Chemistry* **280**: 23397-23407
- Mackintosh SG, Raney KD (2006) DNA unwinding and protein displacement by superfamily 1 and superfamily 2 helicases. *Nucleic Acids Res* **34**: 4154-4159
- Mahony J, McGrath S, Fitzgerald GF, van Sinderen D (2008) Identification and characterization of lactococcal-prophage-carried superinfection exclusion genes. *Applied and Environmental Microbiology* **74**: 6206-6215

- Makarova K, Sorokin A, Novichkov P, Wolf Y, Koonin E (2007) Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol Direct* **2**: 33
- Makarova KS, Aravind L, Grishin NV, Rogozin IB, Koonin EV (2002) A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res* **30**: 482-496
- Makarova KS, Aravind L, Wolf YI, Koonin EV (2011b) Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol Direct* **6**: 38
- 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
- Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, Moineau S, Mojica FJM, Wolf YI, Yakunin AF, van der Oost J, Koonin EV (2011a) Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* **9**: 467-477
- Makarova KS, Wolf YI, Koonin EV (2009) Comprehensive comparative-genomic analysis of type 2 toxin-antitoxin systems and related mobile stress response systems in prokaryotes. *Biol Direct* **4**: 19
- Makarova KS, Wolf YI, van der Oost J, Koonin EV (2009) Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defence against mobile genetic elements. *Biol Direct* **4**: 29
- Malone CD, Hannon GJ (2009) Small RNAs as guardians of the genome. *Cell* **136**: 656-668
- Manica A, Zebec Z, Teichmann D, Schleper C (2011) In vivo activity of CRISPR-mediated virus defence in a hyperthermophilic archaeon. *Mol Microbiol* **80**: 481-491
- Maris C, Dominguez C, Allain FH-T (2005) The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J* **272**: 2118-2131
- Marques JT, Devosse T, Wang D, Zamanian-Daryoush M, Serbinowski P, Hartmann R, Fujita T, Behlke MA, Williams BR (2006) A structural basis for discriminating between self and nonself double-stranded RNAs in mammalian cells. *Nat Biotechnol* **24**: 559-565
- Marraffini LA, Sontheimer EJ (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**: 1843-1845
- Marraffini LA, Sontheimer EJ (2009) Invasive DNA, chopped and in the CRISPR. *Structure* **17**: 786-788
- Marraffini LA, Sontheimer EJ (2010) Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* **463**: 568-571
- Marraffini LA, Sontheimer EJ (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet* **11**: 181-190

- McGrath S, Fitzgerald GF, van Sinderen D (2002) Identification and characterization of phage-resistance genes in temperate lactococcal bacteriophages. *Mol Microbiol* **43**: 509-520
- Medina-Aparicio L, Rebollar-Flores JE, Gallego-Hernández AL, Vázquez A, Olvera L, Gutiérrez-Ríos RM, Calva E, Hernández-Lucas I (2011) The CRISPR/Cas Immune System Is an Operon Regulated by LeuO, H-NS, and Leucine-Responsive Regulatory Protein in *Salmonella enterica* Serovar Typhi. *Journal of Bacteriology* **193**: 2396-2407
- Meister G, Landthaler M, Dorsett Y, Tuschl T (2004) Sequence-specific inhibition of microRNA- and siRNA-induced RNA silencing. *RNA* **10**: 544-550
- Meister G, Tuschl T (2004) Mechanisms of gene silencing by double-stranded RNA. *Nature* **431**: 343-349
- Melderer LV (2010) Toxin-antitoxin systems: why so many, what for? *Curr Opin Microbiol* **13**: 781-785
- Menon AL, Poole FL, Cvetkovic A, Trauger SA, Kalisiak E, Scott JW, Shanmukh S, Praissman J, Jenney FE, Wikoff WR, Apon JV, Siuzdak G, Adams MWW (2009) Novel multiprotein complexes identified in the hyperthermophilic archaeon *Pyrococcus furiosus* by non-denaturing fractionation of the native proteome. *Mol Cell Proteomics* **8**: 735-751
- Mills S, Griffin C, Coffey A, Meijer WC, Hafkamp B, Ross RP (2010) CRISPR analysis of bacteriophage-insensitive mutants (BIMs) of industrial *Streptococcus thermophilus* - implications for starter design. *J Appl Microbiol* **108**: 945-955
- Ming D, Wall ME, Sanbonmatsu KY (2007) Domain motions of Argonaute, the catalytic engine of RNA interference. *BMC Bioinformatics* **8**: 470
- Moazed D (2009) Small RNAs in transcriptional gene silencing and genome defence. *Nature* **457**: 413-420
- Mojica FJM, Díez-Villaseñor C (2010) The on-off switch of CRISPR immunity against phages in *Escherichia coli*. *Mol Microbiol* **77**: 1341-1345
- Mojica FJM, Díez-Villaseñor C, García-Martínez J, Almendros C (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**: 733-740
- Mojica FJM, Díez-Villaseñor C, García-Martínez J, Soria E (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* **60**: 174-182
- Mulepati S, Bailey S (2011) Structural and biochemical analysis of the nuclease domain of the clustered regularly interspaced short palindromic repeat (CRISPR) associated protein 3(CAS3). *The Journal of Biological Chemistry*: 1-18
- Murray NE (2002) 2001 Fred Griffith review lecture. Immigration control of DNA in bacteria: self versus non-self. *Microbiology* **148**: 3-20
- Myong S, Bruno MM, Pyle AM, Ha T (2007) Spring-Loaded Mechanism of DNA Unwinding by Hepatitis C Virus NS3 Helicase. *Science* **317**: 513-516

- Nakata A, Amemura M, Makino K (1989) Unusual nucleotide arrangement with repeated sequences in the *Escherichia coli* K-12 chromosome. *Journal of Bacteriology* **171**: 3553-3556
- Narberhaus F, Vogel J (2009) Regulatory RNAs in prokaryotes: here, there and everywhere. *Mol Microbiol* **74**: 261-269
- Nariya H, Inouye M (2008) MazF, an mRNA interferase, mediates programmed cell death during multicellular *Myxococcus* development. *Cell* **132**: 55-66
- Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, White O, Salzberg SL, Smith HO, Venter JC, Fraser CM (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**: 323-329
- Nowotny M, Yang W (2009) Structural and functional modules in RNA interference. *Curr Opin Struct Biol* **19**: 286-293
- Nozawa T, Furukawa N, Aikawa C, Watanabe T, Haobam B, Kurokawa K, Maruyama F, Nakagawa I (2011) CRISPR Inhibition of Prophage Acquisition in *Streptococcus pyogenes*. *PLoS ONE* **6**: e19543
- Oke M, Carter LG, Johnson KA, Liu H, McMahon SA, Yan X, Kerou M, Weikart ND, Kadi N, Sheikh MA, Schmelz S, Dorward M, Zawadzki M, Cozens C, Falconer H, Powers H, Overton IM, van Niekerk CA, Peng X, Patel P, Garrett RA, Prangishvili D, Botting CH, Coote PJ, Dryden DT, Barton GJ, Schwarz-Linek U, Challis GL, Taylor GL, White MF, Naismith JH (2010) The Scottish Structural Proteomics Facility: targets, methods and outputs. *J Struct Funct Genomics* **11**:167-80
- Palmer KL, Gilmore MS (2010) Multidrug-Resistant Enterococci Lack CRISPR-cas. *MBio* **1** (4)
- Palmer KL, Whiteley M (2011) DMS3-42: The secret to CRISPR-dependent biofilm inhibition in *Pseudomonas aeruginosa*. *Journal of Bacteriology* **193**: 3431-3432
- Paterson S, Vogwill T, Buckling A, Benmayor R, Spiers AJ, Thomson NR, Quail M, Smith F, Walker D, Libberton B, Fenton A, Hall N, Brockhurst MA (2010) Antagonistic coevolution accelerates molecular evolution. *Nature* **464**: 275-278
- Pause A, Sonenberg N (1992) Mutational analysis of a DEAD box RNA helicase: the mammalian translation initiation factor eIF-4A. *EMBO J* **11**: 2643-2654
- Pecota DC, Wood TK (1996) Exclusion of T4 phage by the hok/sok killer locus from plasmid R1. *Journal of Bacteriology* **178**: 2044-2050
- Peng X, Brügger K, Shen B, Chen L, She Q, Garrett RA (2003) Genus-specific protein binding to the large clusters of DNA repeats (short regularly spaced repeats) present in *Sulfolobus* genomes. *Journal of Bacteriology* **185**: 2410-2417
- Perez-Rodriguez R, Haitjema C, Huang Q, Nam KH, Bernardis S, Ke A, Delisa MP (2011) Envelope stress is a trigger of CRISPR RNA-mediated DNA silencing in *Escherichia coli*. *Mol Microbiol* **79**: 584-599
- Phok K, Moisan A, Rinaldi D, Brucato N, Carpousis AJ, Gaspin C, Clouet-d'Orval B (2011) Identification of CRISPR and riboswitch related RNAs among novel

- noncoding RNAs of the euryarchaeon *Pyrococcus abyssi*. *BMC Genomics* **12**: 312
- Pichlmair A, Schulz O, Tan CP, Naslund TI, Liljestrom P, Weber F, Reis E Sousa C (2006) RIG-I-Mediated Antiviral Responses to Single-Stranded RNA Bearing 5'-Phosphates. *Science* **314**: 997-1001
- Polach KJ, Uhlenbeck OC (2002) Cooperative Binding of ATP and RNA Substrates to the DEAD/H Protein DbpA . *Biochemistry* **41**: 3693-3702
- Portillo MC, Gonzalez JM (2009) CRISPR elements in the Thermococcales: evidence for associated horizontal gene transfer in *Pyrococcus furiosus*. *J Appl Genet* **50**: 421-430
- Pougach K, Semenova E, Bogdanova E, Datsenko KA, Djordjevic M, Wanner BL, Severinov K (2010) Transcription, processing and function of CRISPR cassettes in *Escherichia coli*. *Mol Microbiol* **77**: 1367-1379
- 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
- Prangishvili D, Forterre P, Garrett RA (2006) Viruses of the Archaea: a unifying view. *Nat Rev Microbiol* **4**: 837-848
- Prangishvili D, Garrett RA (2005) Viruses of hyperthermophilic Crenarchaea. *Trends Microbiol* **13**: 535-542
- Price EP, Smith H, Huygens F, Giffard PM (2007) High-resolution DNA melt curve analysis of the clustered, regularly interspaced short-palindromic-repeat locus of *Campylobacter jejuni*. *Applied and Environmental Microbiology* **73**: 3431-3436
- Przybilski R, Richter C, Gristwood T, Clulow JS, Vercoe RB, Fineran PC (2011) Csy4 is responsible for CRISPR RNA processing in *Pectobacterium atrosepticum*. *RNA Biology* **8**
- Pul U, Wurm R, Arslan Z, Geissen R, Hofmann N, Wagner R (2010) Identification and characterization of *E. coli* CRISPR-cas promoters and their silencing by H-NS. *Mol Microbiol* **75**: 1495-1512
- Pyle A (2002) Metal ions in the structure and function of RNA. *Journal of Biological Inorganic Chemistry* **7**: 679-690
- Pyle AM (2008) Translocation and Unwinding Mechanisms of RNA and DNA Helicases. *Annu Rev Biophys* **37**: 317-336
- Raivio T (2011) Identifying your enemies--could envelope stress trigger microbial immunity? *Mol Microbiol* **79**: 557-561
- Rajkowitsch L, Chen D, Stampfl S, Semrad K, Waldsich C, Mayer O, Jantsch MF, Konrat R, Bläsi U, Schroeder R (2007) RNA chaperones, RNA annealers and RNA helicases. *RNA Biology* **4**: 118-130
- Redder P, Peng X, Brügger K, Shah SA, Roesch F, Greve B, She Q, Schleper C, Forterre P, Garrett RA, Prangishvili D (2009) Four newly isolated fuselloviruses from extreme geothermal environments reveal unusual morphologies and a possible interviral recombination mechanism. *Environ Microbiol* **11**: 2849-2862

- Roberts RJ, Vincze T, Posfai J, Macelis D (2010) REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res* **38**: D234-236
- Rocak S, Linder P (2004) DEAD-box proteins: the driving forces behind RNA metabolism. *Nat Rev Mol Cell Biol* **5**: 232-241
- Rogers GW (2001) Further Characterization of the Helicase Activity of eIF4A. *Journal of Biological Chemistry* **276**: 12598-12608
- Rogers GW, Komar AA, Merrick WC (2002) eIF4A: the godfather of the DEAD box helicases. *Prog Nucleic Acid Res Mol Biol* **72**: 307-331
- Rousseau C, Gonnet M, Le Romancer M, Nicolas J (2009) CRISPI: a CRISPR interactive database. *Bioinformatics* **25**: 3317-3318
- Sakamoto K, Agari Y, Agari K, Yokoyama S, Kuramitsu S, Shinkai A (2009) X-ray crystal structure of a CRISPR-associated RAMP superfamily protein, Cmr5, from *Thermus thermophilus* HB8. *Proteins* **75**: 528-532
- Samai P, Smith P, Shuman S (2010) Structure of a CRISPR-associated protein Cas2 from *Desulfovibrio vulgaris*. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **66**: 1552-1556
- Sashital DG, Jinek M, Doudna JA (2011) An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. *Nature Structural & Molecular Biology* (ahead of print)
- Schmidt A, Endres S, Rothenfusser S (2011) Pattern recognition of viral nucleic acids by RIG-I-like helicases. *J Mol Med* **89**: 5-12
- Schütz P, Karlberg T, van den Berg S, Collins R, Lehtiö L, Högbom M, Holmberg-Schiavone L, Tempel W, Park H-W, Hammarström M, Moche M, Thorsell A-G, Schüler H (2010) Comparative Structural Analysis of Human DEAD-Box RNA Helicases. *PLoS ONE* **5**: e12791
- Semenova E, Jore MM, Datsenko KA, Semenova A, Westra ER, Wanner B, van der Oost J, Brouns SJJ, Severinov K (2011) Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 10098-103
- Semenova E, Nagornykh M, Pyatnitskiy M, Artamonova II, Severinov K (2009) Analysis of CRISPR system function in plant pathogen *Xanthomonas oryzae*. *FEMS Microbiol Lett* **296**: 110-116
- Sengoku T, Nureki O, Nakamura A, Kobayashi S, Yokoyama S (2006) Structural basis for RNA unwinding by the DEAD-box protein *Drosophila* Vasa. *Cell* **125**: 287-300
- Shabalina SA, Koonin EV (2008) Origins and evolution of eukaryotic RNA interference. *Trends Ecol Evol (Amst)* **23**: 578-587
- Shah SA, Garrett RA (2011) CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems. *Res Microbiol* **162**: 27-38
- Shah SA, Hansen NR, Garrett RA (2009) Distribution of CRISPR spacer matches in viruses and plasmids of crenarchaeal acidothermophiles and implications for their inhibitory mechanism. *Biochem Soc Trans* **37**: 23-28

- Sharanov YS, Zvereva MI, Dontsova OA (2006) Saccharomyces cerevisiae telomerase subunit Est3p binds DNA and RNA and stimulates unwinding of RNA/DNA heteroduplexes. *FEBS Lett* **580**: 4683-4690
- Shaw NN, Arya DP (2008) Recognition of the unique structure of DNA:RNA hybrids. *Biochimie* **90**: 1026-1039
- She Q, Singh RK, Confalonieri F, Zivanovic Y, Allard G, Awayez MJ, Chan-Weiher CC, Clausen IG, Curtis BA, De Moors A, Erauso G, Fletcher C, Gordon PM, Heikamp-de Jong I, Jeffries AC, Kozera CJ, Medina N, Peng X, Thi-Ngoc HP, Redder P, Schenk ME, Theriault C, Tolstrup N, Charlebois RL, Doolittle WF, Duguet M, Gaasterland T, Garrett RA, Ragan MA, Sensen CW, Van der Oost J (2001) The complete genome of the crenarchaeon Sulfolobus solfataricus P2. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 7835-7840
- Shin J-H, Kelman Z (2006) The replicative helicases of bacteria, archaea, and eukarya can unwind RNA-DNA hybrid substrates. *The Journal of Biological Chemistry* **281**: 26914-26921
- Shinkai A, Kira S, Nakagawa N, Kashihara A, Kuramitsu S, Yokoyama S (2007) Transcription activation mediated by a cyclic AMP receptor protein from Thermus thermophilus HB8. *Journal of Bacteriology* **189**: 3891-3901
- Silverman E, Edwalds-Gilbert G, Lin R-J (2003) DExD/H-box proteins and their partners: helping RNA helicases unwind. *Gene* **312**: 1-16
- Singleton MR, Dillingham MS, Wigley DB (2007) Structure and Mechanism of Helicases and Nucleic Acid Translocases. *Annu Rev Biochem* **76**: 23-50
- Singleton MR, Wigley DB (2002) Modularity and specialization in superfamily 1 and 2 helicases. *Journal of Bacteriology* **184**: 1819-1826
- 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
- Siomi H, Siomi MC (2009) On the road to reading the RNA-interference code. *Nature* **457**: 396-404
- Siomi MC, Sato K, Pezic D, Aravin AA (2011) PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol* **12**: 246-258
- Snyder JC, Bateson MM, Lavin M, Young MJ (2010) Use of cellular CRISPR (clusters of regularly interspaced short palindromic repeats) spacer-based microarrays for detection of viruses in environmental samples. *Applied and Environmental Microbiology* **76**: 7251-7258
- Snyder L (1995) Phage-exclusion enzymes: a bonanza of biochemical and cell biology reagents? *Mol Microbiol* **15**: 415-420
- Song J-J (2004) Crystal Structure of Argonaute and Its Implications for RISC Slicer Activity. *Science* **305**: 1434-1437
- Sontheimer EJ (2005) Assembly and function of RNA silencing complexes. *Nat Rev Mol Cell Biol* **6**: 127

- Sontheimer EJ, Carthew RW (2004) Molecular biology. Argonaute journeys into the heart of RISC. *Science* **305**: 1409-1410
- Sontheimer EJ, Marraffini LA (2010) Microbiology: Slicer for DNA. *Nature* **468**: 45
- Sorek R, Kunin V, Hugenholtz P (2008) CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* **6**: 181-186
- Soultanas P, Wigley DB (2000) DNA helicases: 'inching forward'. *Curr Opin Struct Biol* **10**: 124-128
- Soultanas P, Wigley DB (2001) Unwinding the 'Gordian knot' of helicase action. *Trends Biochem Sci* **26**: 47-54
- Stern A, Keren L, Wurtzel O, Amitai G, Sorek R (2010) Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends Genet* **26**: 335-340
- Stern A, Sorek R (2010) The phage-host arms race: Shaping the evolution of microbes. *Bioessays* **33**: 43-51
- Story RM, Li H, Abelson JN (2001) Crystal structure of a DEAD box protein from the hyperthermophile *Methanococcus jannaschii*. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 1465-1470
- Sturino JM, Klaenhammer TR (2004) Bacteriophage defence systems and strategies for lactic acid bacteria. *Adv Appl Microbiol* **56**: 331-378
- Sturino JM, Klaenhammer TR (2004) Antisense RNA targeting of primase interferes with bacteriophage replication in *Streptococcus thermophilus*. *Appl Environ Microbiol* **70**: 1735-1743
- Sturino JM, Klaenhammer TR (2006) Engineered bacteriophage-defence systems in bioprocessing. *Nat Rev Microbiol* **4**: 395-404
- Talavera MA, De La Cruz EM (2005) Equilibrium and Kinetic Analysis of Nucleotide Binding to the DEAD-Box RNA Helicase DbpA †. *Biochemistry* **44**: 959-970
- Tang T-H, Bachellerie J-P, Rozhdestvensky T, Bortolin M-L, Huber H, Drungowski M, Elge T, Brosius J, Hüttenhofer A (2002) Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 7536-7541
- Tang T-H, Polacek N, Zywicki M, Huber H, Brugger K, Garrett R, Bachellerie JP, Hüttenhofer A (2005) Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*. *Mol Microbiol* **55**: 469-481
- Tang TH, Rozhdestvensky TS, d'Orval BC, Bortolin M-L, Huber H, Charpentier B, Branlant C, Bachellerie J-P, Brosius J, Hüttenhofer A (2002) RNomics in Archaea reveals a further link between splicing of archaeal introns and rRNA processing. *Nucleic Acids Res* **30**: 921-930
- Tanner NK, Cordin O, Banroques J, Doère M, Linder P (2003) The Q motif: a newly identified motif in DEAD box helicases may regulate ATP binding and hydrolysis. *Mol Cell* **11**: 127-138

- Tanner NK, Linder P (2001) DExD/H box RNA helicases: from generic motors to specific dissociation functions. *Mol Cell* **8**: 251-262
- Terns MP, Terns RM (2011) CRISPR-based adaptive immune systems. *Curr Opin Microbiol* **14**:321-7
- Thisted T, Gerdes K (1992) Mechanism of post-segregational killing by the hok/sok system of plasmid R1. Sok antisense RNA regulates hok gene expression indirectly through the overlapping mok gene. *J Mol Biol* **223**: 41-54
- Tock MR, Dryden DTF (2005) The biology of restriction and anti-restriction. *Curr Opin Microbiol* **8**: 466-472
- Touchon M, Charpentier S, Clermont O, Rocha EPC, Denamur E, Branger C (2011) CRISPR Distribution within the Escherichia coli Species Is Not Suggestive of Immunity-Associated Diversifying Selection. *Journal of Bacteriology* **193**: 2460-2467
- Touchon M, Rocha EPC (2010) The small, slow and specialized CRISPR and anti-CRISPR of Escherichia and Salmonella. *PLoS ONE* **5**: e11126
- Treangen TJ, Rocha EPC (2011) Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet* **7**: e1001284
- Tuteja N, Tuteja R (2004) Prokaryotic and eukaryotic DNA helicases. Essential molecular motor proteins for cellular machinery. *Eur J Biochem* **271**: 1835-1848
- Tuteja N, Tuteja R (2004) Unraveling DNA helicases. Motif, structure, mechanism and function. *Eur J Biochem* **271**: 1849-1863
- Tyson GW, Banfield JF (2008) Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ Microbiol* **10**: 200-207
- Ussery DW, Binnewies TT, Gouveia-Oliveira R, Jarmer H, Hallin PF (2004) Genome update: DNA repeats in bacterial genomes. *Microbiology* **150**: 3519-3521
- Vale PF, Little TJ (2010) CRISPR-mediated phage resistance and the ghost of coevolution past. *Proc Biol Sci* **277**: 2097-2103
- van der Oost J, Brouns SJJ (2009) RNAi: prokaryotes get in on the act. *Cell* **139**: 863-865
- van der Oost J, Jore MM, Westra ER, Lundgren M, Brouns SJJ (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem Sci* **34**: 401-407
- van der Ploeg JR (2009) Analysis of CRISPR in Streptococcus mutans suggests frequent occurrence of acquired immunity against infection by M102-like bacteriophages. *Microbiology* **155**: 1966-1976
- Vergnaud G, Li Y, Gorgé O, Cui Y, Song Y, Zhou D, Grissa I, Dentovskaya SV, Platonov ME, Rakin A, Balakhonov SV, Neubauer H, Pourcel C, Anisimov AP, Yang R (2007) Analysis of the three Yersinia pestis CRISPR loci provides new tools for phylogenetic studies and possibly for the investigation of ancient DNA. *Adv Exp Med Biol* **603**: 327-338
- Vestergaard G, Shah SA, Bize A, Reitberger W, Reuter M, Phan H, Briegel A, Rachel R, Garrett RA, Prangishvili D (2008) Stygiolobus rod-shaped virus and the interplay

- of crenarchaeal rudiviruses with the CRISPR antiviral system. *Journal of Bacteriology* **190**: 6837-6845
- Viswanathan P, Murphy K, Julien B, Garza AG, Kroos L (2007) Regulation of dev, an operon that includes genes essential for *Myxococcus xanthus* development and CRISPR-associated genes and repeats. *Journal of Bacteriology* **189**: 3738-3750
- Wang R, Preamplume G, Terns MP, Terns RM, Li H (2011) Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage. *Structure* **19**: 257-264
- Waters LS, Storz G (2009) Regulatory RNAs in bacteria. *Cell* **136**: 615-628
- Weir JR, Bonneau F, Hentschel J, Conti E (2010) Structural analysis reveals the characteristic features of Mtr4, a DExH helicase involved in nuclear RNA processing and surveillance. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 12139-12144
- Westra ER, Pul U, Heidrich N, Jore MM, Lundgren M, Stratmann T, Wurm R, Raine A, Mescher M, Van Heereveld L, Mastop M, Wagner EGH, Schnetz K, van der Oost J, Wagner R, Brouns SJJ (2010) H-NS-mediated repression of CRISPR-based immunity in *Escherichia coli* K12 can be relieved by the transcription activator LeuO. *Mol Microbiol* **77**: 1380-1393
- Wiedenheft B, van Duijn E, Bultema J, Waghmare S, Zhou K, Barendregt A, Westphal W, Heck A, Boekema E, Dickman M, Doudna JA (2011) RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 10092-10097
- 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 defence. *Structure* **17**: 904-912
- Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America* **74**: 5088-5090
- Worning P, Jensen LJ, Nelson KE, Brunak S, Ussery DW (2000) Structural analysis of DNA sequence: evidence for lateral gene transfer in *Thermotoga maritima*. *Nucleic Acids Res* **28**: 706-709
- Xue S, Calvin K, Li H (2006) RNA recognition and cleavage by a splicing endonuclease. *Science* **312**: 906-910
- Yang Q, Del Campo M, Lambowitz AM, Jankowsky E (2007) DEAD-box proteins unwind duplexes by local strand separation. *Mol Cell* **28**: 253-263
- Yang Q, Fairman ME, Jankowsky E (2007) DEAD-box-protein-assisted RNA structure conversion towards and against thermodynamic equilibrium values. *J Mol Biol* **368**: 1087-1100
- Yang Q, Jankowsky E (2005) ATP- and ADP-Dependent Modulation of RNA Unwinding and Strand Annealing Activities by the DEAD-Box Protein DED1. *Biochemistry* **44**: 13591-13601

- Yang Q, Jankowsky E (2006) The DEAD-box protein Ded1 unwinds RNA duplexes by a mode distinct from translocating helicases. *Nature Structural & Molecular Biology* **13**: 981-986
- Yoneyama M, Kikuchi M, Matsumoto K, Imaizumi T, Miyagishi M, Taira K, Foy E, Loo Y-M, Gale M, Akira S, Yonehara S, Kato A, Fujita T (2005) Shared and unique functions of the DExD/H-box helicases RIG-I, MDA5, and LGP2 in antiviral innate immunity. *J Immunol* **175**: 2851-2858
- Yoneyama M, Kikuchi M, Natsukawa T, Shinobu N, Imaizumi T, Miyagishi M, Taira K, Akira S, Fujita T (2004) The RNA helicase RIG-I has an essential function in double-stranded RNA-induced innate antiviral responses. *Nat Immunol* **5**: 730-737
- Young RF (2008) Molecular biology. Secret weapon. *Science* **321**: 922-923
- Zegans ME, Wagner JC, Cady KC, Murphy DM, Hammond JH, O'Toole GA (2009) Interaction between bacteriophage DMS3 and host CRISPR region inhibits group behaviors of *Pseudomonas aeruginosa*. *Journal of Bacteriology* **191**: 210-219

APPENDIX I

CRISPR locus constructs

Repeat sequences are highlighted in yellow.

1. CRISPR I (5' to 3')

GATAAAGAGAAAACCGGTTAAGTTCGTTTTTCATGAAGTTGTTTAAAAGTGTGAAAGTTCGAGTCTCAATG
 CGACCGAAACGAATCTTTCTATAATAATTGAACGTTTATAAATGATAGGGTGTATTTCAATTTAACATAA
 AATCCTTGCACCAGAAATTTGTTAAATTAATTACAACATAAATTTGGTCGCATGAAGAGTAAAGGGTAGTC
 ATGAAGATTTATAAGTAAGAAAAGAGAAAAGATAAGGAAGTATAAAAAACACAACAGATTAATCCCAA
 AGGAATTGAAAGGAACTAGCTTATAGTTTAGAAGAAAACAAACAAATAATGATTAATCCCAAAGGAATT
 GAAAGATTTTCAGCTGAAAATTTGAAATCTGTAGATTTGGATGGATTAATCCCAAAGGAATTGAAAGTT
 CCAAATTTGATCTTCTATTGCGTCTTTTATGCTTTTATTAATCCCAAAGGAATTGAAAGATTGTAGT
 CTTTATCAATCCACGTTTCTCTAATCTTG

2. CRISPR II (5' to 3')

CGTTTATAAATGATAGGGTGTATTTCAATTTAACATAAATCCTTGCACCAGAAATTTGTTAAATTAATT
 ACAACTAAAATTTGGTCGCATGAAGAGTAAAGGGTAGTCATGAAGATTTATAAGTAAGAAAAGAGAAAGAA
 AGATAGGAAGTATAAAAAACACAACAGATTAATCCCAAAGGAATTGAAAGGAACTAGCTTATAGTTTAGA
 AGAAAACAAACAAATAATGATTAATCCCAAAGGAATTGAAAGATTTTCAGCTGAAAATTTGAAATCTGT
 AGATTTGGATGGATTAATCCCAAAGGAATTGAAAGTTCCAAAATTGATCTTCTATTGCGTCTTTTATTG
 CTTTTGATTAATCCCAAAGGAATTGAAAGATTGTAGTCTTTATCAATCCACGTTTCTCTAATCTTG

3. CRISPR T7 (5' to 3')

TTGTAATACGACTCACTATAGGGATAGGAAGTATAAAAAACACAACAGATTAATCCCAAAGGAATTGAA
 GGAACTAGCTTATAGTTTLAGAAGAAAACAAACAAATAATGATTAATCCCAAAGGAATTGAAAGATTTTC
 AGCTGAAAATTTGAAATCTGTAGATTTGGATG

APPENDIX II

Multiple sequence alignments

A.2.1 Conserved motifs of Cmr2 family members

The following alignment illustrates conserved motifs between Cmr2 proteins and HD hydrolases. Alignment generated with COBALT, shading with BOXSHADE. Parts of the alignment where no conservation was found are not shown. Protein sequences: *S. solfataricus* Sso1991, *S. islandicus* YP_002836826.1 Cmr2 family, *S. tokodaii* Sto1979, *S. solfataricus* Sso1729, *Hyperthermus butylicus* DSM 5456 HD hydrolase, *Thermoanaerobacter tengcongensis* MB4 hydrolase, *Archaeoglobus fulgidus* DSM 4304 Afu1867, *Nitrococcus mobilis* Nb-231 HD superfamily, *Thermotoga sp.* RQ2 Cmr2 family, *Syntrophus aciditrophicus* SB HD hydrolase, *Sorangium cellulosum* 'So ce 56' HD hydrolase, *Pyrococcus furiosus* DSM 3638 Pfu1129, *Methanocaldococcus jannaschii* DSM 2661 Mj_1672, *Megasphaera micronuciformis* F0359 Csm1, *Fusobacterium nucleatum subsp. vincentii* ATCC 49256 HD superfamily, *Thermotoga petrophila* RKU-1 metal dependent phosphohydrolase, *Thermoplasma volcanium* GSS1 HD superfamily hydrolase.

				HD nuclease motif					
Sso1991_Cmr2	1	MS-----TDD-N-----SREEFLNYKIMALLHDPNPKAWVIT	TSRAH-----NLT	QVLR	S				
S.isl_Cmr2	1	MS-----TDD-N-----SREEFLNYKIMALLHDPNPKAWVIT	TSRAR-----NL	TEQL	S				
Sto1979_Cmr2	1	MS-----KMSLK-----PMSEIIEKEKFAALLHDPNPKPFIFSL	NAF-----NEEK	RISH	S				
Sso1729_Cmr2	1	MR-----RVGSK-----NCSRGGENRKVLLYYLLINK-----	LLK		S				
Hbu_HDhydr	1	MS-----RLH-----MVKLAALLHDPNPKAWVIT	SAF-----GKT	GTVGR	S				
Tte_HDhydr	1	-----M-EKV-----WLLKLFKALLHDPNPKAWVIT	SMNEECIKK	FQLEKGR	S				
Afu1867	1	-----MGQYFRVSGKYRGRENLPREVNKMADNEFWLNKIRAFF	HDPPDKSFE	LKT-----HERR	S				
Nmo_HDhydr	1	-----MTDRL-----WQAKLHARLHDPNPKAWVIT	PAEKALVLL	LRD-----PAGE	S				
Cmr2_TherRQ2	1	-----MSEREEFWKSKITALLHDPNPKAWVIT	PLVKAFD	VKN-----HEDI	S				
Syn_HDhydr	1	-----MTWQIPDTSYWENKFAAYWHDPPDKSFE	LKT-----VTSIQNH	-----EERA	S				
Sce_HDhydr	1	MSERDRDRDRNLAVAHVPVHGR-TGDDEIEHHGTEADRTRFWQKLLQ	LLHDPNPKAWVIT	LRQAG-----GHK	S				
Pfu1129	1	-----M-----VNIK-----EKL	FVYLDHPPDK	KALKEN-----HEER	S				
Mja1672	1	-----MGNCNEYTA-----LKI	GALLHDIGK-----FI	QRASDKPK--SKGH	S				
Csm1_Mmi	1	-----MDQ-----LEARLQKAAALLHDIGK	-----VY	QRSGLG--QGTH	S				
Fnu_HDhydr	1	-----MDE-----KLICLQLGALLHDIGK	-----V	VRAG--LD--SKEH	S				
Tpe_hydr	1	-----MRKGGESLKDREELVVGALLHDIGK	-----V	VRAG--D--DRRH	S				
Tvo1_HDhydr	1	-----MDN-----DEFVLITASLLHDIGK	-----I	QQRKYL-----SEK	S				
Sso1991_Cmr2	44	V--RARK---SH-ERVAKYIINQLFGDINS-----	-----KT-----	VDNADK	S				
S.isl_Cmr2	44	V--RARK---SH-ERVAKYIINQLFGDINS-----	-----KT-----	VDNADK	S				
Sto1979_Cmr2	45	V--KVAKTLISHLDFLGKNELDEISSKIYS-----	-----KKEKGCNSK	VSDADSLASK	S				
Sso1729_Cmr2	31	Y--KELRRFFEN-----FYSYKKS	SGS-----RAK--	ARKKVS	S				
Hbu_HDhydr	37	VLAKVAGKSAPAGELVKVEECEELMKMKTYTIEADAHQLDAAAAVAIT	LDGLRDAPIT	EKLILDEK	S				
Tte_HDhydr	45	IL---IKSKVIEPLLGEDLTEEE--EKL---IEKADTQAYP-----	VNRILP-PVA-----	VKIRGEDV	S				
Afu1867	58	IL-----GELKPSKSLKRIKNI--ADIQASSLQRVDLE-KS	IHKKEKLS	TFD-RIHNT	S				
Nmo_HDhydr	38	VL---HERLFPQGMAG-DLRATV--RKADWWSAADRPQFP-----	RDGKEG-PYAR	WSQVNF	S				
Cmr2_TherRQ2	36	IL---K-TLGIKSRGEE-----DRLASAMDRFPPIPYEKDAKKQIHVS	FD-ET-----	LFVHP	S				
Syn_HDhydr	39	YL---Q-----IFGIDRPNDEFWQADAIAGFERGQVP-----	SHSTDD-TKN--	GAVDFA	S				
Sce_HDhydr	72	LF---Q-ATAGVPLKYVR---PG---PDWAASGADRVPVSS---	PPRPAHV	SVD-WVKN-----	S				
Pfu1129	32	IL---S-SGNIQYSRTDK--VKQ---ADALSSKTQRFIIR-TKENKE	PVIDFLG-R--SSG	KYFHVGY	S				
Mja1672	42	FL----KEKFKNGFLNHLDEKTKDKILEIVKEHHNQ-KI-----	KD-----DL	IGIVRL	S				
Csm1_Mmi	38	FL----KPYEN-----DDSL--VLRVAKYHHAD--MA-----	KRLTKDD-----	DLAYIV	S				
Fnu_HDhydr	38	YL---KNNNLL-----ADRYKEIYDTIDYHHAK-YLS-----	SADLKED-----	SLAYIV	S				
Tpe_hydr	44	FT-----NKV-----KFAVIQDYIHYHHEK-DLL-----	KKSL	LENE-----K	S				
Tvo1_HDhydr	37	FI----KEIKYS-----NKDMERIANLVKHHHDPDKT-----	ELDGRDK-----	KLLKIL	S				

Sso1991_Cmr2 136 LNVNTNTNLNI---LKYQLFYLIYELI--WIDSRYENTP-A-ETRNPTHTIFDHLIYATAAMMN-WIFSLEKE-----
 S.isl Cmr2 136 LNVNTNTNLNI---LKYQLFYLIYELI--WIDSRYENTP-A-ETRNPTHTIFDHLIYATAAMMN-WIFSLEKE-----
 Sto1979_Cmr2 162 LGLTG--VSI---ETYNVYFFYEFLL--WVAKGYTVGE-A-DTRVPTHSIFDHLIYATASIN-WFLG-----
 Sso1729_Cmr2 144 LDKTKIFKLIK--SVYFLIYTIYEPL--WIYFGLPEVE-E-DSRSPFYTFIDHLYASASMIN-WVYVDDSDPKGKKK
 Hbu_HDhydr 183 -----LNLNLYFLLLEPL--WYEVCRACIPLA-DTRTPHTVFDHLYATAAMVN-WLYPGGK-----
 Tte_HDhydr 161 F-----H---WMRI-----HP-A-DTRAPNHSIYDHLVQSTVLS--SALPKPA-----
 Afu1867 195 --EKLKEA-----LLEEFASFAAE-FVNLFP-A-YTLPDHTLFDHADAASALF--GAEIDGK--KPV---
 Nmo_HDhydr 154 DTADDTARLGE--LWRF-----LP-A-DTRIPDHSIWDHLDLVSFAFA--GAFVADANGE-CA---
 Cmr2_TherRQ2 135 --YILEGS--Q--F-----LP-A-DTRIANHSIIDHLDVLSALK--GC-VEGKQVKAS---
 Syn_HDhydr 160 LAEKNIIGLGA--LWHR-----IP-A-DSRFPDHSIWDHNLCSAIS--SCVELGGRAEEVS---
 Sce_HDhydr 200 DDARLQOACLQ--LWR---RLPEEPPPGVAEVVWRHQF-A-DSRAPDHSIWDHLDVLSAISLSSLSFLSRRREGP-VMPW---
 Pfu1129 162 --VKLKEGVKE---FAKSELKLKEEAEKFAEE-FVNLFP-A-DTRFPDHAIWTHLDLTSAL-----SVK--DPT---
 Mja1672 170 IKDFK-GDVSF----EELYQLMQKYTWCIPSVTMMWKAGSLKGLGLDVSFLDHSKTIICALAC-CLYQMYVKENKKK---
 Csm1_Mmi 160 FERKKPDDME----VNEMLQVLEATLSYIPSSSTATNQFA-----DISLYEHVKLTAFAAA-AMYR-WFKAEGIE---
 Fnu_HDhydr 160 LNSFK-ENIN--PEKLAIVLEACCSYFPSSSYVDTE-----DISYDHYVKTAAASA-CFYL-YDKENNIQ---
 Tpe_hydr 159 --SPTPEDVQEIFPTPDDVNFLTYKYFSFIQETRVVEGDM-----DISLYDHLKVTAMLAL-SLYD-YAKENDLK---
 Tvo1_HDhydr 150 IDKLNFNEDPK-YKFFNTLNSILYKDTVAIPSAFYYSKE-----DIPLYHHLKLTAAATL-SLYR-NLKSSDIE---

Zn ribbon

Sso1991_Cmr2 436 TPESRL-KLFELTKFDK-LPQIGEK-----SKRGYEFCTSCGVLPVAVIIMP---KEDEFEKKLIE--L-----GIARD
 S.isl Cmr2 436 TPESRL-KLFELTKFDK-LPQIGEK-----SKRGYEFCTSCGVLPVAVIIMP---KEDEFEKKLIE--L-----GIARD
 Sto1979_Cmr2 449 SPHTQL-SNYNY---IDSIGE-----TKRGFEYCTSCGVLPAMLLP---KDENEYKFFVEEHT-----GKQFN
 Sso1729_Cmr2 479 KVKEKY-KREILPKPNW-FTSFNKEFDYINGKMWYCTVCGNEPAIINFG--KENDDYSATKCEITLRLAYQNRKYVT
 Hbu_HDhydr 505 AEEREYAKSVSIDAGFA-IAESLEEATSKPLKPEFHECSMCGRLPAIVHLADAQVREFAERLGVV-----
 Tte_HDhydr 434 LLELLETEKLLGARKSIREFQLEQK-GRK-----CSLGEFEVLPD-----WEKLR-----
 Afu1867 469 LYIEILTVLNAIESTH--FDKPAWPGAYK-----CTLGHEHLAIGGE-----SREMENVWVKI-----
 Nmo_HDhydr 437 AVYDLVDRVLVAASVRAFDALEQH-GYR-----CSLTAAEAWLTDNPAHLSLPPQQRDQADTLWSRLG-----
 Cmr2_TherRQ2 356 -----RKVSEHAGYK-----ENPGTFYRYLH-----
 Syn_HDhydr 423 TSHSLVQSALAAEKSIKRSVARTTEP-GEK-----CQMCSEFEVLHSHKQWNGEVAGHYADHLKEFWSQLN-----
 Sce_HDhydr 531 LLHHALVSRHGLRKAEEAQAIAAESGEK-----CTLGGLRQALGAGDAGASV-DAQRETARAFWRFRD-----
 Pfu1129 419 LLVKILDSLGERKVTEERFEKSEQLKQWK-----CHVCGENLAIFGD-----MYD-----H-----
 Mja1672 409 ---FEYK-LEGLFE--PYNRGE--NR-----CVLGR-----NEFDKNE-KGYAIRENE-----
 Csm1_Mmi 378 ---YSKEQLEDMDWGSSELSVGD-GMRE-----CNVCH-----TSANPDLLRPPYVLG-----
 Fnu_HDhydr 380 ---YSLEQLKELFDENSSLNKIYS-YTEE-----CTICK-----KAEDESILKKNALDFDE-----
 Tpe_hydr 374 ---YTEKDLEAIFP--DDLNLIQEKGNT-----CKICG-----NRVDR-----LFSIRE-----
 Tvo1_HDhydr 371 ---AKRMKMEIFS-DNAIERDVKFGHSI-----TNMCE-----SCGMDSI-----

Zn ribbon

Sso1991_Cmr2 497 EKDVRSIKNMI--SPGERLCPWCLVKRALGAEPR-----LMRILLG-DLCSVEKIVNEIVSKDVKIE-IPSTSDIASI
 S.isl Cmr2 497 EKDVRSIKNMI--SPGERLCPWCLVKRALGAEPR-----LMRILLG-DLCSVEKIVNEIVSKDVKIK-IPSTSDIASI
 Sto1979_Cmr2 507 DDQIEALKAIL--SPGEKLPWCLIKRAIGVRPE-----FLRVLITSEDLSRFEE-----KDVFI--PSVSHVAFY
 Sso1729_Cmr2 554 DDDLKDLKVMF--KPGEKLPGLCIKRGVYFRLR-----KNLEKVFKS--TDDIAYS-YKKNVIQPRI
 Hbu_HDhydr 572 -----LF--SEGESLCPYLVRRLVSTSDA-----IRRIMNGL-----NLYSLNPRQLYTR-PPSTDELAAM
 Tte_HDhydr 480 ---SKEKGLV--KEKEQLCGVCLAKRLFPKVMK-----EELNLSEEMKFP-----STSEMATI
 Afu1867 522 ---KRWPSSL--RSNERLCAVCAVKRF-----YPKFIETLDFIEFGVGVKVPDIESVSEVAMC
 Nmo_HDhydr 500 ---RKRPTWV--RKGEHLGALATLKRWLPTLFC-----EELKDTLNMMSFRF--VVSTHTMALA
 Cmr2_TherRQ2 377 ---RLTSTKL--AARK-----MARL-----FPGYEDVY-----
 Syn_HDhydr 486 ---PEQKDDVDFKENERLCSVCLIKRLAPHILK-----NSNDHILCNVFRSDNVPSSTEMALH
 Sce_HDhydr 594 ---RNSDD--GAERLCAVCTMKRVLVRAGVATDERGARRVGLTAAWAGPATPLDDVCDRDELVRVFPSTATIAAQ
 Pfu1129 465 ---DNLKSLW--LDEEPLCPMLIKRY-----YPVWIRSK-----TGQKIR-FESVVDVALL
 Mja1672 449 ---SKSERIDYCASFALTDILKNFQ-----MEKTIKFNKAYPIIHLTKNKDNL-----SL
 Csm1_Mmi 422 ---DDGTLACDTGNALA-----Q-----LGQDIL-----NKDVFI-----TS
 Fnu_HDhydr 427 ---EEGIELCSSRGYI-----D-----LGKEVSS-----LYSNNDKFI-----EK
 Tpe_hydr 414 ---GEEIACDFCKEMY-----E-----LGKDLLIKSHVYLAERKNGKFI-----FK
 Tvo1_HDhydr 408 ---YDNAKICIGLEEE-----N-----IGSLLYKYSN-----II-----TD

Sso1991_Cmr2 613 -----LTID-PEEY----WFSEKRRRYFSLFRH-----RITFPSPYALVRADSDYLGDLLE
 S.isl Cmr2 613 -----LTID-PEEY----WFSEKRRRYFSLFRH-----RITFPSPYALVRADSDYLGDLLE
 Sto1979_Cmr2 604 -----LLRENPEEK--GWKDV-----SPYALVRADSDYLGDLLE
 Sso1729_Cmr2 656 -----LPCI-QKEY--GVSDVNNAFIDAL-KGY-----R-----EFYAIVKAADDMTELAR
 Hbu_HDhydr 676 -----YTKAGQKAK---EILGLEAIYDKLLGEF-----GEEFPHMLIAGILASEQRCASHLA
 Tte_HDhydr 593 -----TENFKVEEFPKMSQKITKLEKHRVNP-----SR-----YYAILQMGDGHMGKWLK
 Afu1867 650 LRFNTLLDTLGFDAAKLGDDVKNYETMISELERLSSEVYKMLGEP-----K-----YYAILMMDGDEMCKLLS
 Nmo_HDhydr 610 -----DEG-----KAAEASRLAHALGYKP-----ET-----YYALLMMDGYMGAWLS
 Cmr2_TherRQ 2466 KFDSTT-----DIARN-----NQANKEIEDPA-----KFKNGYIAVLLMMDGDRMGDWML
 Syn_HDhydr 574 -----IDNR-----DK-----YYALLMMDGDMGKLLIN
 Sce_HDhydr 744 GREQGDGAAQRGGLRDRDGKRVPRSKVEALRRRAVESLRRAAKELDRPAGSARRAGDRIPAGSQVALIALDGRISQILL
 Pfu1129 565 -----KKEIDEEK-----VKEVVDFLNAAAYKEIGNPP-----K-----YYAILVMDGDMGKVIS
 Mja1672 548 -----IAFPIIENETE--KRILDFDGLAEKAFERT--GTRK-----IGILKMFVDNLGIEIFT
 Csm1_Mmi 498 -----ATRLWVGDYSVRKDDG--SGCLEFSELAELSGLGKAGIER-----IGVLRADVDNLGAAFV
 Fnu_HDhydr -----
 Tpe_hydr 496 -----EFEKIAEKAPGKK-----IASLLVVDNLGKIFL
 Tvo1_HDhydr 486 -----WRFILQAKYVPLYDDV--RSIKPFSDYFKDESEHK-----M-----LGVLRADVDNMDGLIVA

PALM domain

```

Sso1991_Cmr2 789 GLLVEL-ELINKHK-GFVIYAGGDDLLAMLPE-----VDEVLD-----FVKESRRAFAGVSTG-----
S.isl_Cmr2 789 GLLVEL-ELINKHK-GFVIYAGGDDLLAMLPE-----VDEVLD-----FVKESRRAFAGVSTG-----
Sto1979_Cmr2 750 ILLKEI-QLVNALG-GFVVYAGGDDLLAILPE-----VENALQ-----FVENSRRKIVAGIE-----
Sso1729_Cmr2 771 TLLRDI-KTVEVENYQQIYAGGDDVVALLPE-----IDRLID-----TLIGLERNFVG-----
Hbu_HDhydr 869 TALYDA-EIVAMLG-GFPVYAGGDDVAALAPGYISKGRLENLTKGYATRATHIKDSIRADTGFVFA-----
Tte_HDhydr 690 FALQEVRRIVEETHYKGLVYAGGDDVLALLPE-----VEEVV-----
Afu1867 764 FSVNHVPDVRKNG-TLTIYAGGDDVLVLLPE-----VDTAFDVAATELAMTFSTSWN-----
Nmo_HDhydr 704 FSTVIAREVVEREHIGRVLVYAGGDDLMAMFA-----VSDLIISAMRRLRLAYSGLAPE-----
Cmr2_TherRQ 2562 FS-QFVGKIVDRH-NGMLVYSAGGDDVLALLPE-----ADSVLEECANDIRKFFSGYLEYEIEIENGSDVERFR-----
Syn_HDhydr 658 FSIYGVASII-KDHGRLIYAGGDDVYAFLE-----IGSALPAARKIRDYYSIFRY-----
Sce_HDhydr 879 FAHTIVPWWVEREFSGRLIYAGGDDVLAITAP-----AGEALDLCARLAQLYSAAWVLDTSPGEGPWAWRAKTWTSSTS
Pfu1129 652 FSIREVRSVV-KD-EGLLIYAGGDDVLAILPE-----VDKALEVAYKIRKEFGKSE-----
Mja1672 645 -----YLVYAGGDDTLIVGAW-----DAVWELAKRIRGDFKKFVCYNPYITLSAGIVFVN---PKFE
Csm1_Mmi 614 -----HIVYAGGDDMFIVGAW-----DDLLELAVDIRRAFRRFT--NDKLTFSAGIGLFK---SAFP
Fnu_Hdhydr -----
Tpe_hydr 560 -----MVIYAGGDDLYLVGGW-----NDVLDVAKELREAFGRFTA--NDFMTFSAGYVITD---EKTS
Tvo1_HDhydr 572 -----YIVYAGGDDVTAVGEI-----NKLKFIISDFHNEFNKYFC--KKINISAGVTVVS---PKPF

```

The following alignment illustrates the conserved “GGDEF” domain between Cmr2 family members, diadenylyl cyclases and response regulators. Protein sequences: *S. islandicus* YP_002836826.1 Cmr2 family, *S. solfataricus* Sso1991, *S. tokodaii* Sto1979, *Thermotoga maritima* MSB8 hypothetical protein TM1794, *Archaeoglobus fulgidus* DSM 4304 Afu1867, *Hyperthermus butylicus* DSM 5456 HD hydrolase, *Ignicoccus hospitalis* KIN4/I hypothetical protein Igni_0328, *S. solfataricus* Sso1729, *Thermoanaerobacter tengcongensis* MB4 hydrolase, *Pyrococcus furiosus* DSM 3638 Pfu1129, *Rhodopseudomonas palustris* DX-1 response regulator receiver modulated diguanylate cyclase, *Rhodopseudomonas palustris* BisB5 response regulator PleD, *Bradyrhizobium* sp. ORS278 response regulator PleD, *Afipia* sp. 1NLS2 response regulator receiver modulated diguanylate cyclase, *Nitrobacter hamburgensis* X14 response regulator PleD.

```

Sisl_Cmr2 744 RKKLEKIDVEKEVENS LKYFRTILKEG---RIIVTPAWHVSISSALNRGLLVEL-ELINKHK-GFVIYAGGDDLLAMLPEV
Sso1991_Cmr2 744 RKKLEKIDVEREVENSLKYFRTILKEG---RIIVTPAWHVSISSALNRGLLVEL-ELINKHK-GFVIYAGGDDLLAMLPEV
Sto1979_Cmr2 705 EIEIINDNDIERIISDLKDFLNKNVDRD---RLLLFPSWVHVSISSMNRILKEI-QLVNALG-GFVVYAGGDDLLAILPEV
Tmar1794 528 IEMFQETDD-----LKYAWKLEKEF---KTIQPAYHRGVSRTLIGIFS-QLVGKIVDRHN-GMLVYSGGDDVLAILEPA
Afu1867 731 LERVSD-----ALRVKAKTV---RRLITPAAHSSISRALKNFSVNHVPDVRKGN-GTLIYSGGDDVVLVLLPE
Hbut_HDhydr 836 PSSGCKKGC-----KALPKPE---ATLVTPTYMALSRQMITALYDA-EIVAMLG-GFPVYAGGDDVAALAPG
Igni_0328 719 PKEVYGEND-----SSLP----TVLVTPTYLFQLSYSIMTEALVVK-EIVEKNY-GLLVFAGGDDLLALVPA
Sso1729_Cmr2 745 -----IAKVINDG---NILMSPTYRVALSIAMMITLLRDI-KTVEVENYQQIYAGGDDVVALLPEI
Tthen hydr 654 GEHIKQ-----HAKDNLSSILCKKHPPTPSLHQTLRSKISTFALQEVRRIVEETHYKGLVYAGGDDVLAILEV
Pfu1129 625 RDYV-----EIPPEA---KYYSTPQVHVAISQALANFSTIREVRSVV-KDE-GLLIYAGGDDVLAILEV
Rho cyclase 326 LMIL---DID-----FFKSINDSYG-----HDAGDDVIREFALRIK-KSIRGID--LACRYGGEEFVIVMPE
Rho PleD 326 LMIL---DID-----FFKSINDSYG-----HDAGDDVIREFATRIR-KSIRGID--LACRYGGEEFVIVMPE
Brad PleD 326 LMML---DLD-----YFKSINDTYG-----HDAGDDVIREFAMRVR-KSIRGID--LACRYGGEEFVIVMPE
Afi cyclase 326 LMIL---DID-----FFKSINDTYG-----HDAGDDVIREFATRIR-KSIRGID--LAARYGGEEFVIVMPE
Nitr PleD 326 LMML---DID-----FFKSINDTYG-----HDAGDDVIREFATRIR-KSIRGID--LACRYGGEEFVIVMPE

```

```

Sisl_Cmr2 819 D-----EVLDFVKESRRAFAGVS-----TGR LGNMCLENGFARIN-NAYYPSL-PIVGRSY
Sso1991_Cmr2 819 D-----EVLDFVKESRRAFAGVS-----TGR LGNMCLENGFARIN-NAYYPSL-PIVGRSY
Sto1979_Cmr2 780 E-----NALQFVENSRRKIVAGIE-----DASSYKGFILKIN-NSYFSQL-PLVGRSY
Tmar1794 595 D-----SVLECANDIRKFFSGHL-----EYEIEIESG---SDVERFRSENGVLYHNDKPFAPLMGRAATMSA
Afu1867 795 D-----TAFDVAATELAMTFSTSW-----NGWEMPLPGN---K-----LSA
Hbut_HDhydr 899 Y-ISKG--RLENL--IKGYATRATHIKDSIRADTGFVPALIALY-TRKNYWGLLWAGRCFHRTPIGAVYPAP-VAYGRSY
Igni_0328 780 RSVSRGSGRAEPLGGLEEFLSRELL---EIVKEFYFSPALVWVWLTRLNHWGLLRSPVGFRTD-NFFAPAL-LAYGRSY
Sso1729_Cmr2 802 D-----RLIDTLIGLERNFVG-----ENGFYKVR-QWYIPTF-YPHGRSF
Tthen hydr 722 E-----EVLECAVELQNAFKEVL-----SSEA-----S-----MSA
Pfu1129 682 D-----KALEVAYKIRKEFGKSE-----ENGSLLPGW---K-----LSA
Rho cyclase 382 TDLHVAQ-----MVAERLRRRAIAGEP-----FAVEKCTRRIE-----VTTSI
Rho PleD 382 TDLHVAQ-----MVAERLRRRAIAGEP-----FGIEKGAKRIE-----VTTSI
Brad PleD 382 TDLHVAG-----MVAERLRRSVANEP-----FSVHKCKRID-----VTTSI
Afi cyclase 382 TDLHVAG-----IAERLRRSIANEP-----FSIEKCTKRIE-----VTTSI
Nitr PleD 382 TNLHVAG-----MVAERLRRSIAGEP-----FAVHKAKRID-----VTTSI

```

```

Sisl_Cmr2      868 SVIIAHYA-DPLFFVINDSYNLEEGKEMIRYRVMYNGEYKDAKKDVAIFRYQGLTSVI-----PLSLKRPIV-
Sso1991_Cmr2  868 SVIIAHYA-DPLFFVINDSYNLEEGKEMIRYRVMYNGEYKDAKKDVAIFRYQGLTSVI-----PLSLKRPIV-
Sto1979_Cmr2  824 ILYFSHVK-YPLQLALEESYNLEEGKERVKY----DKYK--KDIVIFKYRNSVSFI-----PLSLIRPYEE
Tmar1794      654 GIAIVHHK-FPLQVALKIARE-AEKRAK--NV-----YGRNAFCVTQVKRSGQMIFA-----GSTW----ETEEED--
Afu1867       826 GLLIVHYK-HPLYDALEKTRELQ-KAK--KL-----GRNAIAVGLLKRSGSYYES-----VVNF----ETLE----
Hbut_HDhydr   972 GIYIVHYR-DFMAAWRSAGDL-EEYVDVIAFTSPHGTTVSKDATFLAYGRVSSIAGVELGAVALPNMKPGAGKEKVTW
Igni_0328     855 GIAIRHYR-DPLAKVFEDASEL-EE-----SAKNVSKKDGVGVSYGRLGA-----RGVALSN---SLGVED---
Sso1729_Cmr2  840 SVRIANIA-DFMTNETQMTTELNRVKK-VKWEFP-NGEEKRS-----KFSAILSTS-----RTSYESVLP-
Tthen_hydr    748 GIVIVHHK-YFLYLALKEVQ-LAQKKAKDERQ-----YNRNAFCLKFKGSGALKEC-----GGKW---ALMDFL--
Pfu1129      713 GILIVHYK-HPLYDALEKARDLNNKAK--NV-----PGKDTLAIGLLKRSGSYIIS-----LVGW----ELIRVfy-
Rho_cyclase   419 GLSTLERKGEPIPDLLKRADTALYRAKHDGRNRVVAAAA-----
Rho_PleD      419 GLSTLERKGEFVRDLLKRADTALYRAKHDGRNRVVAAAA-----
Brad_PleD     419 GISTLEQKGEPIADV MKRADTALYRAKNEGRNRVAIAPVHQPSFLPQAAGRGR-----
Afi_cyclase   419 GISMLEKKSEFVADVLKRADQALYRAKHDGRNRVVADAA-----
Nitr_PleD     419 GLSILERKGEFVADVLKRADIALYRAKHDGRNRVVAQAA-----

```

A.2.2 Multiple sequence alignment of Csa2 orthologues

Protein sequences as annotated from NCBI: *S. solfataricus* P2 Sso1442, *Sulfolobus islandicus* M.14.25 YP_002828994.1, *S. solfataricus* P2 Sso1997, *Acidianus hospitalis* W1 Csa2 YP_004458920.1, *Metallosphaera sedula* DSM 5348 Csa2 YP_001191228.1, *Hyperthermus butylicus* DSM 5456 Hbut_0644 , *Pyrococcus abyssi* GE5 PAB1686, *Methanocaldococcus* sp. FS406-22 Csa2 YP_003458587.1, *S. solfataricus* Sso1399, *Pyrococcus horikoshii* OT3 PH0920, *Sulfolobus tokodaii* Sto0029, *Candidatus Korarchaeum cryptofilum* OPF8 Csa2 YP_001736868.1, *Archaeoglobus fulgidus* DSM 4304 Afu1871, *Pyrococcus furiosus* DSM 3638 PF0642, *Methanocaldococcus jannaschii* DSM 2661 MJ_0381, *Pyrobaculum aerophilum* str. IM2 PAE0210.

```

Sso1442      1 M-I-----SGSVRFLVNLLESNLNGVESIG-NLTKHRTAPVVLK-TSTGYLVRYVPVISGEALAHAYQASLVD--IAKKE
Sisl_Csa2   1 M-I-----SGSGRFLVNLLESNLNGVESIG-NLTKHRTAPVVLK-TSTGYLVRYVPVISGEALAHAYQASLVD--IAKKE
Sso1997     1 M-I-----GGSGRFLVNLLESNLNGVESIG-NLTKHRTAPVVLK-TSTGYLVRYVPVISGEALAHAYQASLVD--IAKKE
A.hosp Csa2 1 M-I-----SGSARFLINVESNLNGVESVG-NLTKHRTAPVVVK-TSTGYLIRYVPVISGESLAHAYQASLVD--IAKSM
Msed_DevR family 1 M-----ISGSVRFVLVNLLESNLNGVESV-NLSRHRTPAPIVTRKSTGEYVIRYVPVISGESLAHAYQALVVE--IAEKM
Hbut0644    1 MPV-----FFSLSARILVNLEALNMAESV-NVVRHRRAPVVLK-TDNGFVLRYVPVISGESLAHHYQKLLAD--IAIQR
Pab1686     1 M-M-----FLSVGVRFEANVEALNMVETAG-NYTKHRRVPYLVE-EDGKLTIVYVPAISGESLAHAYQELLVK--EALRM
M/coccus_DevR fam 1 --M-----FISIGVRFANVEALNMVETAG-NYSKHRRVPYIE-EDGKLTIVYVPAISGESLGHAYQELLVK--ESKAL
Sso1399     1 MQLVNNMWISFSVRYLVNVEDLNNVESAG-NYVRHRRAPLVFK-DKDSYTVTYVPAVSSEMIAHGYQMNLVE--LAIQR
PH0920     1 M-M-----FLSVGIRFEANVEALNMVETAG-NYTKHRRVPYLIE-ENGKLTIVYVPAISGESLAHAYQELHVN--EALSA
Sto0029    1 M---VKMKWVSFSARYLVNVEDLNNVESAG-NYVRHRRAPIIVK-EGNTYTVTYVPAVSSEMIAHGYQMNLVE--IAIER
CKc_DevR family 1 ----MADPFVSVRGRVLINVEALNMTEVSG-NYVKHRRVPVIMP--E-TYATYFVPSVSGESIAHGYQQVLAE--EASGK
Afu1871    1 M-VVSDVVFVSVRGRVMLNVEAMNMTEVSG-NYVKHRRVPVPLP--DAKYTTYFVPAISGESIAHGFQEVLAE--VGKKN
Pfu0642    1 M-M-----YVRISGRIRLNAHSLNAQGGG-TNYIEITKTKVTVR-TENGWTVVEVPAITGNMLKHHWFVGFVD--YFKTT
Mja0381    1 --M-----FLRISGRVRLNSHSLNAQGGG-TNYVEITKAKVSIK-NDDRWEILEVPAISGNMVKHHWFVSVFD--FFRET
Pae0219    1 M-V-----YVRVTARVEVQVSAISGLGAI-NYNQVATARI LHN-G----ALYEVVITGNALKHHAVYAVEAYQALGG

```

```

Sso1442      69 GLPVGSLSSQYEFIKFSTDEALKIEGIEKPKDYNDAR----RFEVEVMLKDVIADVGGFMYAGGAP---VRRTRIRIKLGY
Sisl_Csa2   69 GLPVGNLSSQYEFIKFSTDEALKIEGIEKPKDYNDAR----RFEVEVMLKDTIADVGGFMYAGNAP---VRRTRIRIKLGY
Sso1997     69 GLPVGNLSSQYEFIKFSTDEALKIEGIEKPKDYNDAR----RFEVEVMLKDVIADVGGFMYAGSAP---VRRTRIRIKLGY
A.hosp Csa2 69 NLPVGLYSSQYEFIKYSSDEVLKEEGISAPSSSDNDR----RFEVEVLLKDIVSDVGGFMYAGKYP---VRRTRIRIKFGY
Msed_DevR family 70 GLPVTHRTKQGELIK-FANDDV-LKEENIASPKDEKAR--RFEVDVMLKDVVADVGGFMYAGKNP---VRRTRIRIKLGY
Hbut0644    72 GLPVCACSQGVFLK-HANDDV-FKKYDGDIGAKNPKFTGTDAAEYVVKNCVVEDVGGFLYTDKTV---KRTSFRFVGY
Pab1686     71 NLPVCDCHRGFEFYK-SMNKVH-LQKKISPIP-NDPR----KIEEAIIRKCVVEDVGGFLYAEKPP---VRRSSAFQVSY
M/coccus_DevR 70 NLPVCDCEKFEFFK-SMNKNY-LKKKINPVPKDDK----KIEEAIKSCVIEDVGGFLYAEKPP---VRRSSAFQVSY
Sso1399     77 NLPVDSLAKGILIKRGSDDK--HEGKTCDEKGS-----YELCVINEDIVEDVAGFMNPNKLV---KRTSNVAFSY
PH0920     71 GLPVCDCCRGEFYK-SMNKIH-LEKKVSP-IP-DDPK----EIEEAIKACVVEDVGGFLYAEKPP---VRRSSAFQVSY
Sto0029    74 NLPVEELAKQGILIKRAGDSV--HK-TGCGDKNGSD----YELCVIEEDIVEDVAGFLNPKLV---KRTSNVAFSY
CKc_DevR family 71 GLPVCKLCSKGYFLK-STNDAV-FKESFGVNPPEGES----EFERAVIKGCVVEDVGGFLYAPARGGKNVKTSNFFVGY
Afu1871    75 GLKVCCKLCEKGIFLK-STNENV-FKESFSSDPPKDDF----EFKTVIENCIVEDVGGFLYAPRAGG-NVKRTSNFYTYG
Pfu0642    72 PYGN-LTERALRYN-GTRFQGQETTATKANGATVQL----NDEATI KKLADAVHGFLAPKTVGR---RFLVSVKASF
Mja0381    71 DYKDN-LTERALRYN-GARFGQ-ETKAKKADGSEVEL----KDESEI LKNFADADYHGFLAPKTVGR---RFLVSVKTSF
Pae0219    69 NMLNE-LCKRGIGLR---GFTV-DSTLKNPKVPTDECE-----ALKDFCNDLHGFLSPQEEKP---VKKRDSLVKISE

```

```

Sso1442      142 MIPALRGDEIPAQ-LEAQFHVRFNSKNPVSGSQ-----AIFNVEVSSALTLTFSFELDEDLIAVPSTFGEKVKGE-
Sisl_Csa2   142 MIPALRGDEIPAQ-LEAQFHVRFNSKNPVSGSQ-----AIFNVEVSSALTLTFSFELDEDLIAVPSTFGEKVKGE-
Sso1997     142 MIPALRGDEIPAQ-LEAQFHVRFNSKNPVKGSQ-----AIFNVEVSSALTLTFSFELDEDLIAVPSTFGEKVKGE-
A.hosp Csa2 142 MIPALTGEELPAQ-LEAQFHVRYSSK-VEERQ-----AIFNVEVSSALTLTFSLDDDLIAVPSTIGNEVEGE-
Msed_DevR   143 MIPSLKTDEIPAQ-LEAQFHVRYSVVSKDKQ-----AIYNVEVGSALTLTVSFLDDGLIGVPSNPGKADKDE-
Hbut0644    146 MVEALDALEAGAAATEAQFHVRYSPGAKQEQ-----AIYYVEIGSAVVVFSFALDSAGVGSARMENEGASSRN
Pab1686     141 ALP-IKSMALFAT-AEPQLHARHAQIDTSSKK---G-NVSEQMIYYVETGTALMGFVFNLDLDGIGVSAITSEPVLGE-
M/coccus_DevR 141 ALP-IKSIAYAT-TEPQLHARHAQTGEGKKE---G--VAEQMIYYVETGTAVMGFTFNLDLDAIGISSLTNKAVVDE-
Sso1399     146 MVP-AIDAVKAST-ISSQFHVRYANKELMDYK---NEN--IQSLYNIETASASVVLTYGLVNVSVGTQNYPVKEVDK-
PH0920     141 ALP-VKSVALFAT-SEPQLHARHAQIDASSKK---G-NVSEQMIYYVETGTALMGFVFNLDLDAVIGSAITSKPILDD-
Sto0029    142 MIP-ALDAVKASA-VTSQFHVRYATKEMI DKYE---KENKNIQSLYNVETASASVVLTYGLVNLSNIGVTQNYPVKEVDK-
CKc_DevR   145 MIP-TRESLESAV-IEPQLHRYALGTPFVEE---GARAGQMYYIELSSAATLTFSFELDTKYLGKATFSMENVGGT-
Afu1871    148 MIP-VRESIEGAV-IEPQLHSRYALGTPFVEG---G---QGQMIYYVELSSAVTLTFSFELDTTRYIGRTTFSYEKAGTE-
Pfu0642    142 ILP-TEDFIKEVEGERLITAIKHNRVVDDEKGAIGSSKEGTAQMLFSREYATGLMGFSIVLDLGLVGPQGLPVKFEVND-
Mja0381    140 ILP-TEDFIKEVD-ERLVYAVKHNRVDIDEKGAIGSSKEGTAQMLFNREYATGLMGFSIVLDLGLVGPQSSP-----
Pae0219    133 AVVLEEGNLKAV---AKFAVQHNRRVVPPTVNV--KQKEGEGMMLFKQEGYGTGLAFALRMLAHIGNPLFDECNAEFQ-

```

Appendix

Sso1442 209 -----EELERQKARRVDSAIKALYSLLAG-NFGGKRSFLEFSMKLMSLVVTKTD-FFF-MPEPAHDDSYIKTT-IMRLGK
 Sisl_Csa2 209 -----EELERQKARRVDSAIKALYSLLAG-NFGGKRSFLEFSMKLMSLVVTKTD-FFF-IPEPGHDDSYIKTT-VMRLEK
 Sso1997 209 -----EELERQKARRVDSAIKALYSLLAG-NFGGKRSFLEFSMKLMSLVVTKTD-FFF-IPEPGHDDSYIKTT-VMRLEK
 A.hosp_Csa2 208 -----EELEGQKTRVKAATKSLYSILITG-NFGGKRSFLEFSMKLMSLVVTKTD-FFF-IPEPGHTDDSYIKVS-VERLNLK
 Msed_DevR 211 LL-----RIRGARVEASVRATYHLLTG-NFGGKRSFLEFSMKLMSLVVTKTD-FFFVVEVPG-HSDDYIKLS-HERAER
 Hbut0644 215 EY----LSLEDRLKRVAAFDALAAALGGMAWGAKTSFQFHHWKILSLVASVSOPLFNFVSPG-HDKNMARET-VERACA
 Pab1686 213 -----E-EIKKRREAAALMALFRMLSSAQFGAKLSFFFVGGITELVVSVTE-HFFVVTSP-IYEGYAEKT-EKRLEV
 M/coccus_DevR 212 -----D----GIKKRREASLKAIFRMLSSQFGAKLSFFFVGNIMEVAIAITE-HFFSVTSP-IYDNMMEKT-EKRLEK
 Sso1399 218 -----KKDREKAALDALMLTITQPLFGAKLTFKFIIVEIEALFVSASE-KFNLPPV--TGDIKKYI-DLVNST
 PH0920 213 -----N----EIKKRREVSLLKALFRMLSSQFGAKLSFFFVGGITELIVTVTE-HFFVVTSP-IYDDYIERT-KRRLNI
 Sto0029 216 -----KKDREIASLDALMLTITQPLFGAKLTFKFIIVEIEALVLSVSE-KFNLPPV--NGDENYDL-NLVKST
 CKc_DevR 218 -----VVDGDERKKRIGAAALDALSKFMIEMMFQAKTFFLEFVIEWSVVIASD-DVWTVSP-FSKNMIERA-EEKVVK
 Afu1871 218 -----V----GKRSEIRINAALALKKFIEFAFGAKKTFFLEFVMEWDSLVAVSD-DVWTVSP-YTAGYIDNA-RKKKEK
 Pfu0642 221 PRPNVIDPNERKARIESALKALIPMLSG-YIGANLARSFFVFKVEELVAIASE-GPIPALVHGFEYEDYVEAN-RSIIKN
 Mja0381 212 ---NPVIEDDERKARIVSALKALIPMLSG-YIGANLARSFFVFKLEEMIAVSE-KPIPALVHGFEYEDYVEVS-KNVVEN
 Pae0219 207 -----SDERKRRAKASVLLALPLITG--AGSKQAALFIVAVREVLAVSE-KMPNLIHAVYPDYICETSIDTVGAY

Sso1442 280 AKGV-LNGNLAKAYVINNEGIEVGEV-TVLSTVEDLVVKLEEE-----
 Sisl_Csa2 280 AKSV-LNGNVAKVYVINNEGIEVGEGA-TVLSSVEDLVFKLKEK-----
 Sso1997 280 AKSV-LNGNVAKVYVINNEGIEVGEGA-TVLSSVEDLVAKLKEK-----
 A.hosp_Csa2 279 AKSI-FNSKNVEVFTINNENIEVPSNV-KTLSSAEDLIDELIKSKK-----
 Msed_DevR 280 AKSILMGKKVKTFAINRE-GLDTGKAE--VKSNPVEEVVEALLKEVKG-----
 Hbut0644 290 MTSVVKGFKASIVYNGEGLMEPEGCANNVSEKVGSYLEAIRRAKEETLELLRGKS
 Pab1686 281 L-K--S-FNEDYFYTKTS---EDKLPE---EVLKEVTEYIREKEYI-----
 M/coccus_DevR 280 I-A--DTFREEIKFMTD---GEKTPE---ECLAEMINYVKDKNII-----
 Sso1399 283 TDSF-----AKILNKRPPVVKYLYLKE--EKGNVNTPIDAFVM-----
 PH0920 281 L-K--N-FGEEIFTTIAK---EDRVAE---EALKEATDYLNKGVF-----
 Sto0029 281 ADSF-----SSALEIDRPKIVFYVKG--VKGSLSNPVEVFKSVRG-----
 CKc_DevR family 290 V-S--Y--NTKLFKYT----GGAGFE---EVVIEAMNEAKRRAGVS-----
 Afu1871 287 V-N--F--NTKLFVYP----EGGSFE---EVVVEATEEAKERAGK-----
 Pfu0642 298 ARA--LGFNIEVFTYNVD---LGEDIEATKVSSVEELVANLVKMVGKKE-----
 Mja0381 286 AKK--LGFEIEDFGYNVD---FGE----SVSSVEELLSKIEKL-----
 Pae0219 276 LNGIGDTAKFYYYGRCNADKVGKINFKKVGSLHELIDAVINDVQSWIR-----

A.2.3 Multiple sequence alignment of Cas6 orthologues

Protein sequences as annotated from NCBI: *S. solfataricus* Sso2004, *S. islandicus* Cas6 YP_002836977.1, *Sulfolobus islandicus* REY15A ADX84853.1, *S. solfataricus* Sso1437, *Metallosphaera sedula* DSM 5348 Msed_1137, *Metallosphaera cuprina* Ar-4 Mcup_1148, *Acidianus hospitalis* W1 YP_004458915.1, *Sulfolobus tokodaii* str. 7 Sto2642, *Sulfolobus acidocaldarius* DSM 639 Saci_1864, *Staphylothermus marinus* F1 Smar_0329, *S. islandicus* YP_002828978.1. Residues highlighted in red are predicted by Phyre to be located in the central beft between the two ferredoxin--like domains (figure 4.14). The G-rich loop is colored in blue.

Sso2004	1	M-----PLIF---KIGYNVIFLQDVLVTPSSKVLKYLIIQSGKLIPLSKDLITSRDKYKPIFI
Sisl Cas6	1	M-----PLIF---KIGYNVIFLQDVLVTPSSKVLKYLIIQSGKLIPLSKDLITSRDKYKPIFI
SislREY Cas6	1	M-----PLIF---KIGYNVIFLQDVLVTPSSKVLKYLIIQSGRLPLSKDLITSRDKYKPIFI
Sso1437	1	M-----PLIF---KIGYNVIFLQDVLVTPSSKVLKYLIIQSGKLLPSLNNLITSRDKYKPIFI
Msed1137	1	MHRNLALHAPPKCSYYPRLTCQFMQLM---KMTFNVVPLHDVVLPLSSKVLKYLIVLSQQVLPFLEELVRSKDKQKPLFI
Mcup1148	1	M-ANVELNV-----MQIV---RLNFSVRLRDVVLPMPTSKVVKYLIIQSGKLVLPFVKDLVESKRKQKPLFI
AhospCas6	1	M-----LA---LVKTTYNTPLTDVVLVPLSSKVLKYLIIQSGKLPFSLANLVKSRDKQKPLFI
Sto2642	1	M-----VEFFSE--KIV---KVEFSAVPESDVILPLSSKVLKYLIIQSGKLLPSLSSLVQSGMKNKPLFI
Saci1864	1	M-----TLIVS---AEIDVILPKHDVILPLLSKVAKFIILSKNQ---QIGELIGSKKPKKELSI
Smar0329	1	M-----LITDIIGSRLLYSAS-PRYY---KAHVILEVKGAVLPYPTGKVVKTLINAE--PGLEDVFSNNYKPIAI
Sisl Cas6	1	M-----MIVG---EVFVKPENDTIIIP--FSSKIGKSLILD-----PKSVSI
Sso2004	56	SHLG--FNQRRIFQT----NGNLKTIITKGSRLSSIIAFST---QANVLSEVADEGI--FETVYKGFHIMIESIET-VEV
Sisl Cas6	56	SHLG--FNQRRIFQT----NGNLKTIITKGSRLSSIIAFST---QANVLSEVADEGI--FETVYKGFHIMIESIET-VEV
SislREY Cas6	56	SHLG--FNQRRIFQT----NGNLKSIITKGSRLSSIIAFST---QSNVLPEVADESI--FETVYKGFHIMIESIET-VEV
Sso1437	56	SHLG--LNQRRIFQT----NGNLKTIISRGSKLSSTIAFST---QVNVLPPEL-DEGV--FETVYKGFHITIESVET-VEV
Msed1137	78	SNLA--LDGKRLYS-----RGEPIITVAKARTLTSVTFPF--SKEAFNVG--GGR--VKTVYGEYEISLKEVSV-LD-
Mcup1148	63	SNLG--LNGKRLYSTHEMIRRGDVIKVKAFTKMSASVSFPM---MGEIMNMG--GGR--VSTPYGDFEILLESINV-FN-
AhospCas6	56	SNLG--YGDVRLISD-----GSEVIKINANSRLKATLSFPF---LDGIQNEI-TEGV--YETPYKGFSLDLSIET-VDI
Sto2642	61	SNLGG--NGFRLLFST-----GKPVSVKAGELINFFISFPY---YDGFTEL-SSGS--FETGYKGFTELEQLLEV-IEL
Saci1864	53	SPLSS--NGRFLYAE---NDGKLLRAMRGEKLFHTFSVATSEVNEKTFDLD---GD--VSTPYGDFEYVLLRTIYI---
Smar0329	70	STLAKRVNNKYLVLWKKK--GSDIVLKVDPGDTVEFWGFTEDIASKMIEALTSLDGLKLFNKWSLLEYNIESYKLPKPKP
Sisl Cas6	36	SPL-R-YKGYLV--KNA-SVPTYLEVIGGNVYSFEIGGDEKNVYSALINL-DSKYL--FNMFVKVIDVKVHEIEV-TSI
Sso2004	123	EKLKEEVEKHMNDNIRVRFVSPTELLSSKVLVLPSSLSERYKKIHAGYSTLPVSGLIVAYAYNVYCNLIGK---KE--VEV
Sisl Cas6	123	EKLKEEVEKHMNDNIRVRFVSPTELLSSKVLVLPSSLSERYKKIDAGYSTLPVSGLIVAYAYNVYCNLIGK---KE--VEV
SislREY Cas6	123	EKLKEEVEKHMNDNIRVRFVSPTELLSSKVLVLPSSLSERYKKIDAGYSTLPVSGLIVAYAYNVYCNLIGK---KE--VEV
Sso1437	122	EKLKEEVEKHMNDNIRVRFVSPTELLSSKVLVLPSSLSERYKRVNAGYSTLPVSGLIVAYAYNVYCNLIGK---KE--VEV
Msed1137	141	----ETPSTSTRGNLRVSLFPLALLCSKIYLPFLREKYRKKIGFSLIPTPLVAVGYRQYLALLGKTD-SYE--NDI
Mcup1148	132	----GFNSDVEGKLNKVRIVTPALLSSKIYLPFL-ERYRKAKVGLSLIPSPGLVASAYRTYLGGLGSTE-NEE--EDL
AhospCas6	122	KSLKN-VNNYENANIYVKFLTPTELLSSKILLPSSLSAKYKQVNSGFSLPISIGLIAYAYRNYAAILGNTN-GEE--YAS
Sto2642	126	SSIKGVSE----GNFYVKFVTPALLSSKVLVLPSSLSKEKYKNNVPGYSLIPVSGLVVSYAYRVYRALYGNTS-NME--LDS
Saci1864	118	NQLKDIRHEIKERNVNLRFESPTLLSNKYMVPVVF--KPKKVRSMNRLIPQPSLTFSLANLWNSIADERERIVKGDLEW
Smar0329	149	EEPLDYRLDDAIAVKVEFRTEALL-----LDYKKTTRYKR-----FLPTPGNVFSY---NIGDLLRLTRD-KE--YIE
Sisl Cas6	107	P--KNF-----ELEIMTEALI-----VSEYVKEKKV-----FTNKSEYVF---FNNVTDVTGLNRGDEK--LNE
Sso2004	197	RAFKFGILSNALSRIIGYDLHPVTVAIGEDSKGNLRKARVMGW---IEFD-I-PDERLKRRALNYLLTSSYLIGIGRSSR
Sisl Cas6	197	RAFKFGILSNALSRIIGYDLHPVTVAIGEDSKGNLRKARVMGW---IEFD-I-PDERLKRRALNYLLTSSYLIGIGRSSR
SislREY Cas6	197	RAFKFGILSNALSRIIGYDLHPVTVAIGEDSKGNLRKARVMGW---IEFD-I-PDKRLKRRVLKYLLTSSYLIGIGRSSR
Sso1437	196	RAFKFGVSNALSRIIGYDLHPVTVAIGEDSKGNLRKARVMGW---IEFD-I-PDEKLKRRALRYLLASSYLIGIGRSSR
Msed1137	214	KTFKLLVMANALSRVGYRLYPETVVIGEDKGRRLRLTRGVKGV---IEFD-I-VG-KLKESAAYLEVASFLIGIGRSSR
Mcup1148	204	KSFKLVVNLVNGLSKVVDFELKPVTVIIGEDDKGRLRKSRVVEGW---IMFD-V-TG-KLKRAVAKYLSVASYLGVGKRSR
AhospCas6	198	RAFKLGVLIINAFTKIVGFNLRPKTVIIGRDSKRLRETRGTIGW---IEFD-V-VHDKFKRLAIEYLLIASYLGIGRSSR
Sto2642	199	KSFRLGVLSNLSRVIGYKLPVTVIIGNDNKGRLRTRSGFVGV---MEFD-I-PYKLLKKAISKYLLIASYLGIGRSSR
Saci1864	196	TPYIGRIADVAFAEIGYSLRPVTVIIGKDNQRIQARQFVGV---VKYEVINVNPRYLETFERLEGLAKIFIGIGRSSR
Smar0329	212	VVILVNALLNETYTVLE-TVKPVKYVYGN-----KSLPGIIGYAKYIMIDWD--LLAETKAKHLLLENILLHASIMIGIGRSSR
Sisl Cas6	160	VIYFSAQLLWEEPSVM-----KYTSVRYDD-----KLVIILTEK----LRYS-IK-GE--DEILVVKVLENAIARIGIGRSSR
Sso2004	271	GIGFGEIRLEFRKIEE--KEG-----
Sisl Cas6	271	GIGFGEIRLEFRKIEE--KEGKYTSSDSKG
SislREY Cas6	271	GIGFGEIRLEFRKIEE--KEGKYTSSDFKG
Sso1437	270	GIGFGEIKLEFIKREE--NH-----
Msed1137	287	GIGLGEVHFVKMVERGE---NSH-----
Mcup1148	277	GIGLGEIKLDLVDRSKVEQEGSN-----
AhospCas6	272	GIGLGEIKFELKRRKD-----
Sto2642	273	GIGLGEVVVKIKS-----
Saci1864	272	GIGLGRVVKVE-----
Smar0329	284	ANGFCHVTIKVIQSNE-----
Sisl Cas6	222	RNGFGVVRVKGVDVSW-----SR-

