
THE UNIVERSITY
*of* EDINBURGH

# Structural Investigation of the Archaeal Replicative Machinery by Electron Microscopy and Digital Image Processing


Giuseppe Cannone

**Doctor of Philosphy**
**The University of Edinburgh**

**2014**

*To My Mom...*

*"Science is like sex, it may give some practical results, but that's not why we do it."*

*Adapted from* **RP Feynman**

*I'm not afraid*
*they'll stamp me flat.*
*Grass stamped flat*
*soon becomes a path.*
**Blaga Dimitrova**

# Contents

# Acknowledgements

First and the foremost, I am thankful to my supervisor Dr. Laura Spagnolo for giving me this great opportunity, for believe in me, for priceless advices, for supporting me throughout my PhD and helping me grow as a scientist.

I am grateful to Ken and Noreen Murray, The founder of the Darwin Trust of Edinburgh, for funding my PhD and without whom this thesis would not have been possible.

I would like to thank my second supervisor Prof. Bettina Böttcher, whose assistance and help during these years is priceless. I also would like to thank Dr. Katharina Hipp for introducing me Mausi, Steve Mitchell for help and assistance, Dr. Chris Kennaway for help and advices.

I am thankful to James Parker for being a good friend and lab mate, for the music, the chats, the usage of English and spell checking.

I am thankful to Roberta Carloni and Nicola Festuccia for giving me an house when I first arrived, for all these years of friendship. I will never forget our dinners, party and all the fun time spent together.

I am thankful to Cristina Furlan for her friendship, loyalty, support, for providing always wise advices and for all the nice time together and for the awesome food!

I am thankful to Flavia Scialpi for her friendship, help and spell checking the thesis.

I would like to thank David, Carolina and Christopher for their friendship and time spent together, the documentaries and the no sense talks about philosophy and science.

I am thankful to Emanuele, Maria, Matteo e Claudia for their friendship and support during these years.

I am thankful to you all and to all the people who made these incredible years as nice as they would have been back home. I could not have wished any better.

Last but not the least, ringrazio la mia famiglia per l'amore e il supporto che mi hanno dato in ogni giorno della mia esistenza ringrazio Mamma per aver coltivato l'amore di una famiglia unita e felice nonostante tutte le difficolta'. Ringrazio Luigi e Angela per l'affetto e le gioie che mi hanno regalato in questi anni. Davidino, per l'affetto, i sorrisi e gli abbracci che mi regala.

# Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, either in whole or in part, in any previous application for a degree. Except where otherwise acknowledged, the work presented is entirely my own.

Giuseppe Cannone

# Lay Summary

Cells store their genetic information into a 'biological hard drive' called nucleus. These information are crucial for a proper functioning of cells. The information is stored in a sequence of nucleotides (A, C, G, T) that form DNA. Once a cell divides into two daughter cells, this information must be copied so that each daughter cell receives a complete identical genetic information.

During my PhD, I investigated how cells duplicate their DNA. This process is called DNA replication. Replication start from origins of replication and goes through until it terminates at the end of the genome. Replication factories, scattered throughout the genome copy the DNA at an impressive rated of 3000 nucleotides per minutes and an error rate of 1 nucleotide in a billion. In the process of DNA replication the two strands of DNA are separated, and new daughter strands are built using the base pairing rules which specify that A pairs with T and G pairs with C.

Understanding the mechanism and regulation of DNA replication is important not only for the basic biology of the process but also for medicine. Genetic disease, like cancer, is caused by mutations in DNA replication and repair. Proteins, which are involved in the replication of the DNA double strand, may be at high levels in those cells that are dividing quickly. Thus, any drug that is able to inhibit DNA replication, and thus slowing down cell division, may be used for treating cancer.

# Abstract

Previous studies suggest a degree of homology between eukaryotic replication, transcription and translation proteins and archaeal ones. Hence, Archaea are considered a simplified model for understanding the complex molecular machinery involved in eukaryotic DNA metabolism. DNA replication in eukaryotic cells is widely studied. In recent years, DNA replication studies expanded on the archaeal DNA replication machinery.

*P. abyssi* was the first archaeon whose genome was fully sequenced. Genome sequencing and comparative genomics have highlighted an MCM-like protein in *P. abyssi*. In this study, I report the biochemical and structural characterisation of *Pab*MCM. *Pab*MCM is explored as model for understanding more complex eukaryotic MCM proteins and unravelling the biochemical mechanism by which MCM proteins release their helicase activity.

The crenarchaeon *Sulfolobus solfataricus* possesses a simplified toolset for DNA replication compared to Eukaryotes. In particular, *S. solfataricus* has a subset of the eukaryotic Okazaki fragment maturation factors, among which there are a heterotrimeric DNA sliding clamp, (the proliferating cell nuclear antigen, PCNA), the DNA polymerase B1 (PolB1), the flap endonuclease (Fen1) and the ATP-dependent DNA ligase I (LigI). PCNA functions as a scaffold with each subunit having a specific binding affinity for each of the factors involved in Okazaki fragment maturation. Here, the 3D reconstruction of PCNA in complex with the Okazaki fragment maturation proteins PolB1, LigI and Fen1 is reported.

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| **ARS** | Autonomously Replicating Sequences |
| **Cdc45** | Cell division cycle 45 |
| **Cdc6** | Cell division cycle 6 |
| **CDK** | Cyclin-Dependent protein Kinase |
| **Cdt1** | Chromatin licensing and DNA replication factor 1 |
| **CV** | Column Volumes |
| **DDK** | Dbf4-Dependent protein Kinase 4 |
| **Dpb11** | DNA Polymerase B 11 |
| **dsDNA** | DNA double stand |
| **FEG** | Field Emission Gun |
| **GINS** | Go-Ichi-Ni-San, 5-1-2-3 in Japanese |
| **GuCl** | Guanidinium Chloride |
| **IDCL** | Inter Domain Connecting Loop |
| **LB** | Luria–Bertani broth |
| **MCM2-7** | Mini-Chromosome Maintenance 2-7 |
| **MCM2-7** | Minichromosome maintenance 2-7 |
| **MSA** | Multivariate Stastical Analysis |

**MSA** Multivariate Statical Analysis

**MW** Molecular Weight

**OD** Optical density

**ORBs** Origin Recognition Boxes

**Orc1-6** Origin Recognition Complex 1-6

**ORF** Open Reading Frame

**PCNA** Proliferating Cell Nuclear Antigen

**PIP** PCNA Interacting peptide

**pre-RC** pre-Replicative Complex

**RPA** Replication Protein A

**SANS** Small Angle neutron scattering

**SAXS** Small Angle X–ray scattering

**SD** Shine-Dalgarno

**Sld2** Synthetic lethality with Dpb11 2

**Sld3** Synthetic lethality with Dpb11 3

**SNR** Signal–to–Noise Ratio

**SOB** Super Optimal Broth

**SOC** Super Optimal Broth with Catabolite repression

**ssDNA** DNA single strand

**TAE** Tris-Acetate EDTA

**TB** Terrific Broth

# Chapter 1

# Introduction

## 1.1 DNA replication: an overview

DNA replication is the biological mechanism by which dividing cells replicate the bulk of their DNA. DNA replication is a semi–conservative process [Kornberg and Baker, 1992]. Replication occurs with the two strands of the DNA duplex being copied by base pairing with complementary nucleotides. The result of this process is two DNA duplexes identical to each other and to the parental DNA duplex [Kornberg and Baker, 1992]. During DNA replication, several proteins establish multiple interactions in an astonishing coordination of enzymatic activities aimed to replicate the genetic information. Large multi protein complexes are recruited onto DNA forming the 'replisomes' [Kornberg and Baker, 1992].

DNA replication is carried out in three stages: initiation, elongation and termination [Kornberg and Baker, 1992]. Initiation begins at specific sequences, called origins of replication, with the recruitment and activation of replication initiator proteins, resulting in the formation of the pre–initiation complex (pre–IC), in eukaryotes [Bell and Dutta, 2002] and archaea [Barry and Bell, 2006] or the pre–priming complex in bacteria [Mott and Berger, 2007]. The initiation phase terminates with the recruitment of more replicative factors, such as polymerases, resulting in the assembly of the replication fork [Kornberg and Baker, 1992]. Upon activation of the replication fork, replication proceeds either unidirectionally (e.g. viruses) or bidirectionally (e.g. bacteria, eukaryotes and archaea) from the origin to the sites

of termination [Kornberg and Baker, 1992]. DNA is replicated by enzymes known as DNA–dependent DNA polymerases mostly known as DNA polymerases. DNA polymerases synthesise the complementary strand of DNA in 5'–3' polarity, using an RNA fragment as primer and the single–stranded DNA as template. RNA primers are needed since DNA polymerases require a free 3'–OH to perform the synthesis of the complementary strand [Kornberg and Baker, 1992]. Owing to the anti–parallel configuration of the DNA double strand, during DNA replication, one daughter strand (the leading strand) is synthesised continuously, while the second daughter strand (lagging strand) is synthesised discontinuously by short RNA–primed DNA fragments, also known as Okazaki fragments [Kornberg and Baker, 1992]. Okazaki fragments are then processed, resulting in the removal of the RNA primer and ligation of two, adjacent, Okazaki fragments. As a consequence, the new daughter strand is covalently ligated into one continuous complementary strand [Kornberg and Baker, 1992].

DNA replication terminates in bacteria when the replication fork runs into specific regions of the chromosomal DNA, called the termination sites. The termination step is mediated by association of specific DNA–binding proteins with the termination sites, which arrest the coming replication fork at the specific point [Kornberg and Baker, 1992].

The mechanism of DNA replication has been extensively studied and is relatively well understood in simple organisms like *E. coli* [Johnson and O'Donnell, 2005]. On the other hand, eukaryotic DNA replication mechanisms are far more complex than in prokaryotes. Although the core replicase components are structurally and functionally quite similar, a larger network of protein–protein and DNA–protein interactions is required for assembly, propagation and regulation of the eukaryotic replication fork [Takeda and Dutta, 2005].

Archaea, which represent a separate domain of unicellular organisms, have eukaryotic–like replisomes. These organisms can therefore be used to understand the eukaryotic replisomes [Edgell and Doolittle, 1997]. Additionally, investigating the archaeal replisomes could lead to discovery important biochemical processes unique in these organisms.

## 1.2 Archaea as model to study DNA replication

Carl Woese first reported Archaea as a separate domain of life [Woese and Fox, 1977]. Since then, a growing interest in these fascinating organisms and their unique evolutionary lineage has come into being [Barry and Bell, 2006; Lindås and Bernander, 2013]. Archaea belong to a domain of prokaryotic organisms which thrive in the most diverse range of environments, including those with high temperature (hyperthermophiles), high osmotic pressure (halophiles) and extreme pH (acidophiles and alkalophiles), often in combination with anaerobic growth conditions [Pikuta et al., 2007]. Moreover, non-extremophilic archaea are globally distributed in both marine and terrestrial environments [Robertson et al., 2005]. Currently, phyla recognized within the domain Archaea are: Euryarchaeota [Woese et al., 1990]; Crenarchaeota [Woese et al., 1990] and Thaumarchaeota [Brochier-Armanet et al., 2008]. Several new archaea lineages have been discovered so far, but their phylum status awaits confirmation [Lindås and Bernander, 2013].

Over the past 20 years, genome sequencing of a number of archaea has shed light on an interesting similarity between eukaryal and archaeal proteins involved in replication, transcription and translation [Edgell and Doolittle, 1997; Olsen and Woese, 1997; Lindås and Bernander, 2013]. Archaea also present bacterial–like features, including the lack of a nucleus, the presence of single circular chromosomes and the organization of a large fraction of genes into operons. Introns have not been found in any archaeal gene [Olsen and Woese, 1997; Barry and Bell, 2006; Edgell and Doolittle, 1997]. Comparisons of the amino acid sequences and structural evidence of the component of archaeal replisomes suggest that archaeal replicative proteins are more similar to eukaryal than analogous bacterial proteins (Table 1.1) [Edgell and Doolittle, 1997]. Interestingly, many studies have been documenting an eukaryotic–like cell cycle with characteristic checkpoint–like inhibition of the genome segregation and cell division. The cell cycle of *S. solfataricus* is the best characterised to date [Lindås and Bernander, 2013].

Features such as simplicity of the archaeal replicative machinery and thermostability of hyperthermophilic archaeal proteins, have made hyperthermophilic archaea

an appealing model for studying DNA replication.

### 1.2.1 The euryarchaeon *Pyrococcus abyssi*

The euryarchaeon *Pyrococcus abyssi* (*Pab*) was first isolated from deep–sea hydrothermal vent in the North Fiji basin, at 2000 m depth [Erauso et al., 1993]. Vital cells are highly motile cocci of 0.8–2 $\mu$m in width with a polar tuft of flagella. *P. abyssi* grows at temperatures between 67°C and 102°C under atmospheric pressure, with an optium temperature of 96°C and a doubling time of 33 minutes. *P. abyssi* can grow under hydrostatic pressures of 20–40 MPa.

*P. abyssi* shows a bacterial–like mode of replication [Myllykallio et al., 2000]. Like bacteria, *P. abyssi* possesses a single origin of replication, high replication rate ($\sim$20 kb/min) and bidirectionality of the replication fork [Myllykallio et al., 2000]. Nonetheless, only eukaryal–like proteins are encoded by *P. abyssi*'s genome (Table 1.1)[Cohen et al., 2003].

### 1.2.2 The crenarchaeon *Sulfolobus solfataricus*

*Sulfolobus solfataricus* (*Sso*) was first isolated in the Solfatara volcano (Pisciarelli, Naples, Italy) [Zillig et al., 1980]. *S. solfataricus* thrives in volcanic springs with optimal growth occurring at pH 2–3 and temperatures of 75–80 °C [Huber and Prangishvili, 2006]. *S. solfataricus* is acidophilic and thermophilic with irregularly shaped cells ($\sim$2 $\mu$m) and flagellar. *S. solfataricus* has been extensively used as model organism for research on mechanisms of DNA replication, cell cycle, transcription, RNA processing, and translation [Pfeifer F, 1994]. *S. solfataricus* possesses the most eukaryotic–like mode of replication among archaea (Table 1.1) [Duggin and Bell, 2006]. Similarly to the multi origin organisation of eukaryotic chromosomes, in *S. solfataricus* three origins (*OriC1,OriC2,OriC3)* of replication were discovered to be active in initiation of DNA replication [Robinson et al., 2004; Lundgren et al., 2004].

**Table 1.1: DNA replication eukarytic–like proteins from *P. abyssi* and *S. Solfataricus*.**

| | | *P. abyssi* | *S. Solfataricus* |
|---|---|---|---|
| **Complex** | **Protein** | **Gene name (AN)[a]** | |
| Cdc6/Orc1 | Cdc6/Orc1 | cdc6 (PAB2265) | cdc6–1 (SSO0257)<br>cdc6–2 (SSO2184)<br>cdc6–3 (SSO0771) |
| MCM | Mini–Chromosome maintenance | cdc21 (PAB2373) | mcm (SSO0774) |
| | DNA polymerase I | polI (PAB1128) | dpo1 (SSO0552) |
| DNA pol II | DNA polymerase II small subunit<br>DNA polymerase II large subunit | polB (PAB2266)<br>polC (PAB2404) | dpo2 (SSO1459) |
| | DNA polymerase III | polB (PAB2266) | dpo3 (SSO0081) |
| | RNAse HII | rnhB (PAB0352) | rnhB (SSO2384) |
| DNA primase | DNA Primase small subunit<br>DNA Primase big subunit | priS (PAB2235)<br>priL (PAB2236) | priS (SSO1048)<br>priL (SSO0557) |
| RP–A | Replicative protein A14<br>Replicative protein A29<br>Replicative protein A39 | rpa2 (PAB2164)<br>rpa3 (PAB2165)<br>rpa1 (PAB2163) | ssb (SSO2364) |
| PCNA | PCNA | pcna (PAB1465) | pcna1 (SSO0397)<br>pcna2 (SSO1047)<br>pcna3 (SSO0405) |
| | Flap–endonuclease 1 | fen (PAB1877) | fen (SSO0179) |
| | DNA Ligase I | lig (PAB2002) | lig (SSO0189) |
| RF–C | Replication factor C small subunit<br>Replication factor C large subunit | rfcS (PAB0068)<br>rfcL (PAB0069) | rfcS (SSSO0768)<br>rfcL (SSO0769) |

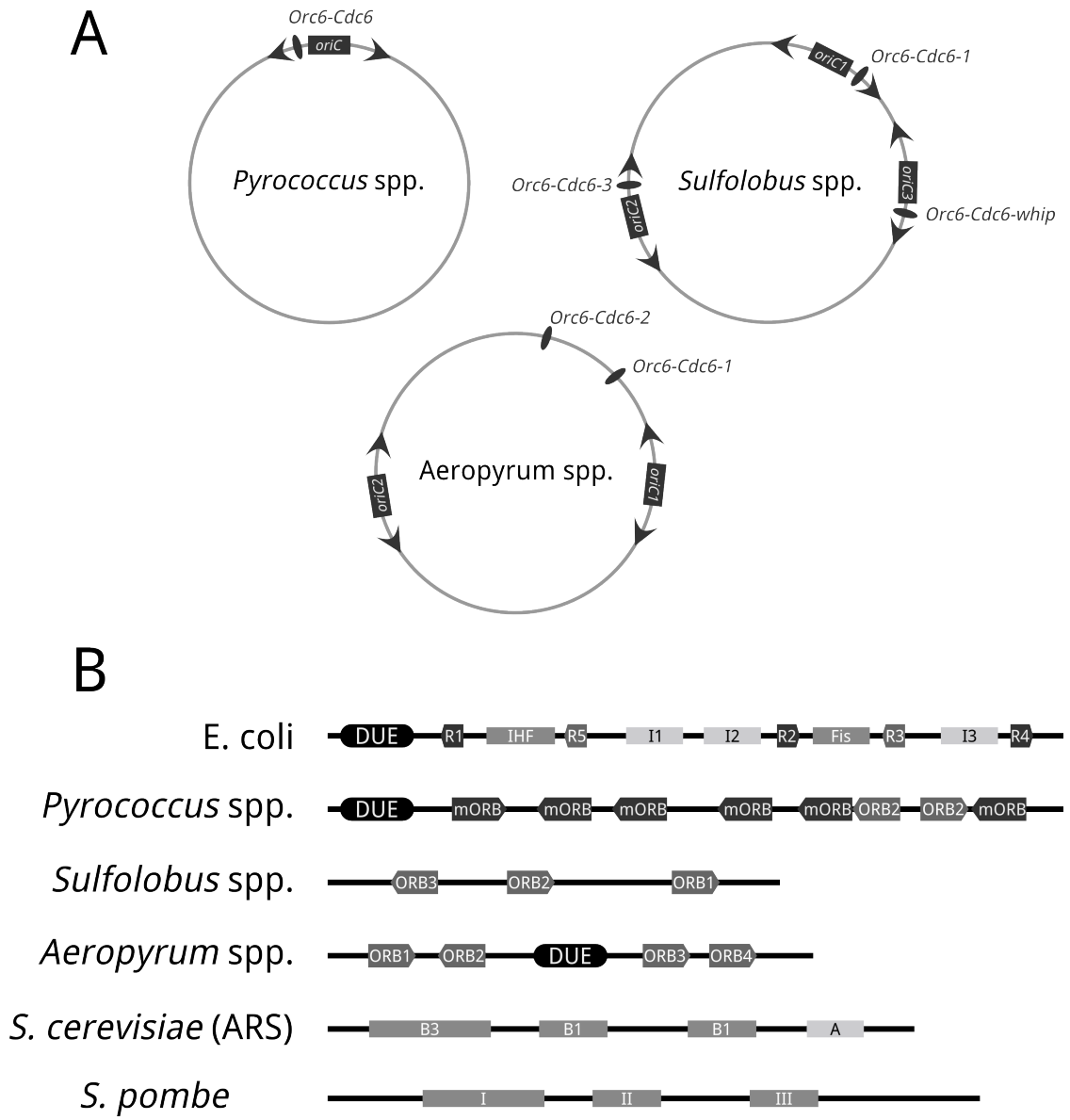



**Figure 1.1: Origin of replication architecture of some well studied organisms**. (A) DNA replication origins in three well–studied archaeal models. Origins are indicated with rectangular boxes while initiator proteins are shown in dark ovals. (B) Functional elements in some well-studied replication origins. The AT-rich DNA–unwinding elements (DUE) represent the sites where unwinding occurs during formation of the 'open complex' [Leonard and Méchali, 2013]. *E. coli oriC* contains five classes of DnaA box consensus [Mott and Berger, 2007]. 'R' and 'I' site are Dna box consensus for DnaA initiator protein [Mott and Berger, 2007]. IHF and Fis DNA box consensus for IHF and Fis DNA-bending proteins [Mott and Berger, 2007; Leonard and Méchali, 2013]. Archaea *oriC* contains three classes of functional elements. Origin recognition box (ORB) and mini origin recognition box (mORB) are the box consensus for Orc6/Cdc6 initiator protein. DUEs are not yet well defined in *Sulfolobus* [Leonard and Méchali, 2013]. *S. cerevisiae* autonomously replicating sequences (ARS) is composed of four functional elements (A, B1, B2, B3) [Gilbert, 2001]. *S. pombe*, origin of replication consist of multiple AT–rich elements that contribute partially to origin activity [Gilbert, 2001]. In both, budding and fission yeast the initiator protein is a multi protein complex formed of ORC, which binds to the origin and recruits Cdc6 and Cdt1 [Bell and Dutta, 2002].

## 1.3 Archaeal origins of replication

An origin of replication is a specific DNA sequence at which replication is initiated on a chromosome [Kornberg and Baker, 1992].

Bacterial chromosomes have a single origin of replication, termed *oriC*, which directs the formation of the pre–priming complex at the origin of replication [Mott and Berger, 2007]. In *E. coli*, *oriC* is approximately 250–bp in length and contains five classes of repeat sequences, referred to as DnaA boxes, which consist of sequence–specific binding sites for the initiator protein DnaA and the architectural factors IHF and Fis (Figure 1.1 B) [Mott and Berger, 2007].

In eukaryotic species, the sequences required for initiation vary significantly among different organisms (Figure 1.1 B) [reviewed in Gilbert, 2001]. For example, some systems require specific DNA sequences, whereas in others any DNA sequence can promote initiation of DNA replication [Bell and Dutta, 2002]. Evidence suggests that in higher eukaryotes, origins are defined by a variety of other DNA binding proteins rather than sequence–specific DNA recognition elements [Barry and Bell, 2006].

*P. abyssi* was the first Archaeon whose chromosomal replication origin (*oriC*) was identified *in vivo* [Matsunaga et al., 2001]. Subsequently, more origins of replication were discovered [Barry and Bell, 2006; Wu et al., 2014]. Archaea use a single or multiple origin(s) of replication to replicate the bulk of their DNA (Figure 1.1 A) [Kelman and Kelman, 2004; Robinson and Bell, 2005; Wu et al., 2014]. The archaeal origin(s) of replication consists of an origin(s) region (*oriC*) and one or more initiator genes called Orc6/Cdc6 [Barry et al., 2007]. The origin(s) region consists of a AT-rich DNA–unwinding elements (DUE) flanked by several conserved repeated motifs known as origin recognition boxes (ORBs) [Kelman and Kelman, 2003; Leonard and Méchali, 2013]. Although the number, orientation, and spacing of ORBs vary among archaea, ORB motifs can be classified in two major classes: a long ORB motif (22–35 bp) and a shorter ORB motif (12–13 bp), termed mini–ORB. Both consist of inverted repeats with dyad symmetry. Mini–ORBs are often found in multiple (7–15) direct repeats analogous to some bacterial origins [Leonard and Méchali,

2013].

## 1.4 DNA replication: initiation phase

In 1963, Jacob et al.'s seminal paper proposed the 'replicon model', an intriguing mechanism by which a *trans*–acting initiator protein would bind a *cis*–acting replicator DNA sequence to initiate the replication of DNA in bacteria. Over the past 50 years, the 'replicon model' has proved to be extremely accurate, and the cis–acting DNA sequence is now know as the origin of the replication, whilst DnaA, ORC–Cdc6–Cdt1 and Orc6/Cdc6 are known as *trans*–acting initiator proteins in bacteria, eukaryotes and archaea, respectively [Bell and Dutta, 2002; Mott and Berger, 2007; Barry and Bell, 2006].

The replication process begins in an ordered fashion with the recruitment of initiator proteins at the origin [Bell and Dutta, 2002]. In eukarya and archaea, initiation results in the assembly of a nucleoprotein complex termed the pre–replicative complex (pre–RC) [Bell and Dutta, 2002; Barry and Bell, 2006], while in bacteria this is termed the pre–priming complex [Mott and Berger, 2007]. Initiation is mandatory to assembling two bidirectional replication forks at the origins of replication [Bell and Dutta, 2002; Mott and Berger, 2007; Kornberg and Baker, 1992].

Careful coordination of initiation of DNA replication, with the events occurring during the cell cycle, is vitally important in order to maintain genome stability [Blow and Dutta, 2005]. This coordination is achieved in two steps: origin licensing and origin activation [Bell and Dutta, 2002; Blow and Dutta, 2005].

### 1.4.1 Origin licensing

Origin licensing marks the origin(s) to prevent premature activation of chromosomal replication, while activation of the origin(s) ensures that chromosomal replication is timed with the cell cycle.

The bacterial model of pre–priming complex formation is provided in Figure 1.2. Origin licensing entails regulation of DnaA proteins and DNA methylation at the level of interspersed 'GATC' sites [Mott and Berger, 2007]. Further events ensure

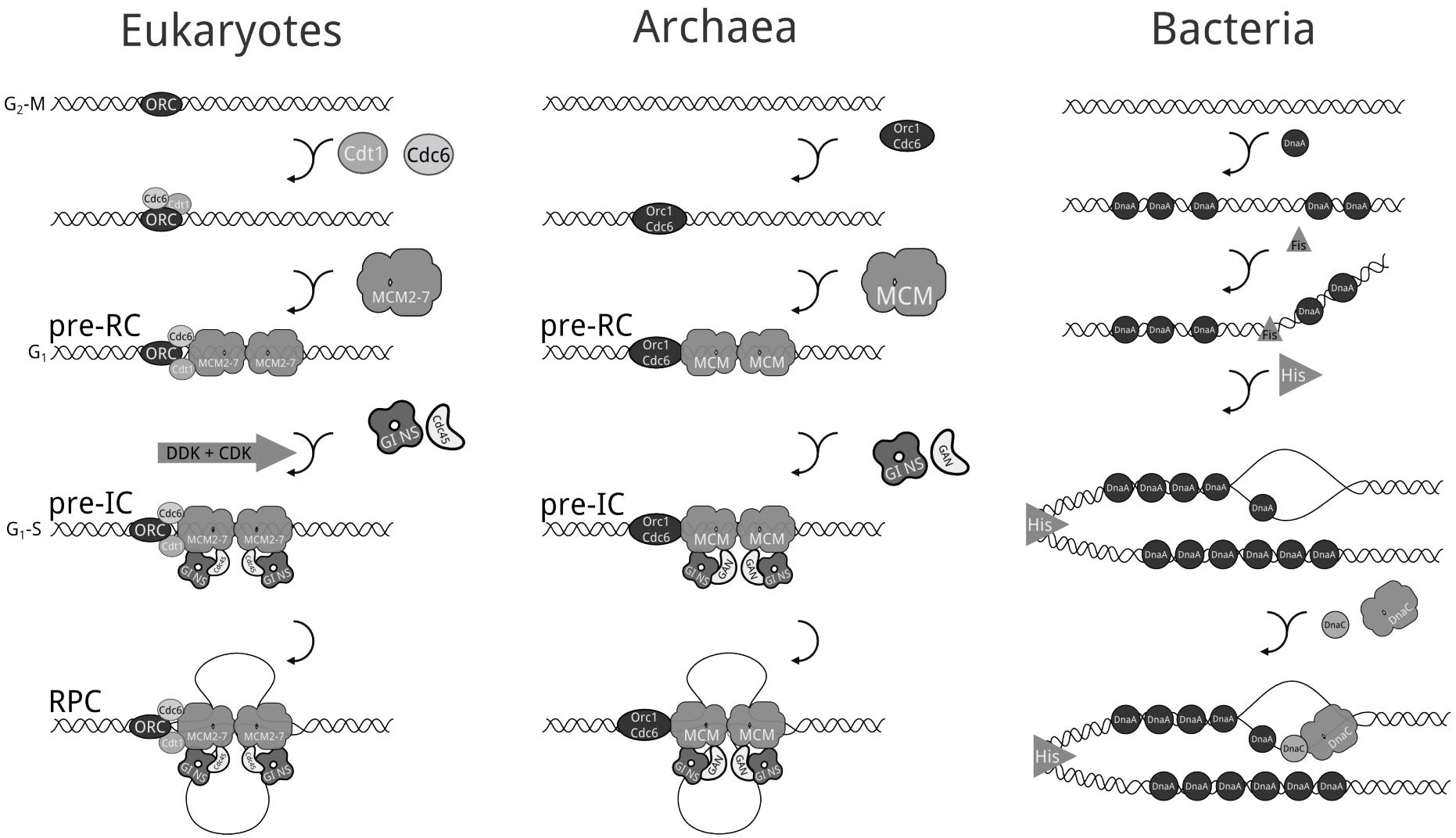



**Figure 1.2: Schematic representation of origin licensing and activation in eukaryotes, archaea and eubacteria.** (Eukaryotes) Components of the pre–RC are recruited at the origin in a stepwise manner. ORC first binds to the origin. Next, OCR recruits Cdc6 and Cdt1, which are both required for MCM2–7 recruitment and loading. The origin of replication is now licensed. Two kinases activate the pre–RC to the pre–IC, which lead to recruitment of more factors and further origin unwinding. (Archaea) Similarly to eukaryotes, ORC first binds to the origin. Next, OCR recruits Cdc6, which are both required for MCM recruitment and loading. The origin of replication is now licensed. Kinases may activate the pre–RC to the pre–IC, which lead to recruitment of more factors and further origin unwinding. (Bacteria) DnaA binds the origin, further recruitment of architectural factor (Fis and His) and DnaA oligomerisation, induce drammatic toloplogical changes which results in oring unwinding. DnaC loades the replicative helicase DnaB.

origin unwinding within the DUE consensus and loading of the replicative helicase DnaB. Activation is stimulated after the replicative fork is assembled [Kornberg and Baker, 1992].

In eukaryotes, origin licensing is achieved by assembling the pre–replicative complex (pre–RC) in an ATP–dependent manner. The eukaryotic pre–RC is a nucleoprotein complex composed of ORC, Cdc6, Cdt1 and mini–chromosome maintenance protein complex (MCM2-7) [Bell and Dutta, 2002; Blow and Dutta, 2005]. ORC is recruited at the origins during late $G_2$–phase/early M phase, which in turn recruits Cdc6, Cdt1 and the MCM2-7 complex during $G_1$ phase. The principal function of ORC, Cdc6 and Cdt1 is to load the MCM2-7 protein complex, hence forming the pre–RC [Bell and Dutta, 2002; Takeda and Dutta, 2005].

Archaeal origin(s) licensing occurs in an eukaryotic–like fashion (Figure 1.2) [Barry et al., 2007; Leonard and Méchali, 2013; Wu et al., 2014]. Most archaeal genome sequenced to date, with only few exceptions, contain at least one gene homologous of both ORC and Cdc6 [Barry et al., 2007]. Owing to the homology with both ORC and Cdc6, archaeal Orc/Cdc6 genes are usually annotated as Orc1/Cdc6–x [Barry et al., 2007].

*In vivo* studies with *Pab*Orc1/Cdc6 proteins shown that *Pab*Orc1/Cdc6 binds to the *oriC* throughout the cell cycle, similarly to the eukaryotic ORC [Matsunaga et al., 2001; Bell and Dutta, 2002]. Biochemical observation of other archaeal Orc1/Cdc6 proteins have shown specific origin binding to ORB elements as well as unwinding an *oriC in vitro* [Matsunaga et al., 2010; Robinson et al., 2004].

Bioinformatic analyses, of *S. solfataricus* genome, revealed a new initiator protein called Whip (Winged–Helix Initiator Protein, Whip). Whip show a sequence–specific binding to the *oriC3* of origin of *S. solfataricus* and hence it has been suggested as candidate for the missing archaeal Cdt1 [Robinson and Bell, 2007].

Structural studies of Orc1/Cdc6 from *P. aerophylum* and Orc2/Cdc6 *A. pernix* revealed that Orc1/Cdc6 proteins consist of three structural domains, two N–terminal $AAA^+$ modules and a C–terminal winged helix (WH) fold [Liu et al., 2000; Singleton et al., 2004]. This structural evidence confirms that Orc1/Cdc6 proteins belong to the family of $AAA^+$ ATPases. In the same group, appear proteins termed

clamp loaders [Takeda and Dutta, 2005], ATP-fuelled molecular machines which open, load and close ring–shaped molecules onto DNA [Jeruzalmi et al., 2002]. It was suggested that ORC and Cdc6 may have a role as clamp loaders since ATP binding and hydrolysis is required for Cdc6–dependent loading of MCM2-7 complex [Takeda and Dutta, 2005]. Nevertheless, EM studies of yeast pre–RC complex showed that the ORC–Cdc6 complex forms a ring-shaped structure [Speck et al., 2005] with ORC arranged as Orc1–4–5–2–6–3 and Cdc6 closing the gap between Orc1 and Orc–3 and hence forming an heptamer ring with dimensions similar, to those of the ring-shaped MCM helicase [Chen et al., 2008; Sun et al., 2012]. Recent cryo–EM studies revealed the architecture of the yeast pre–RC complex. The overall structure forms a double heptameric ring, Cdt1 forms extensive contacts with MCM2-5-7 while the ORC–Cdc6 complex showed structural similarity to the replication factor C clamp loader, hence suggesting a conserved mechanism of action [Sun et al., 2013].

The crystal structures of *Ape*Orc1/Cdc6•DNA [Gaudier et al., 2007] and the heterodimeric *Sso*Orc1/Cdc6–1–3•DNA bound [Dueber et al., 2007] showed that DNA binding is mediated by a C–terminal WD domain that inserts deeply into the major and minor grooves, widening them both, resulting in unwinding of the duplex at the binding site [Gaudier et al., 2007; Dueber et al., 2007]. Additionally, DNA contacts are also established though the N–terminal $AAA^+$ domain. Interestingly, the crystal structure of DnaA shows a striking structural similarity with the archaeal Cdc6/Orc1 [Erzberger et al., 2002]. DnaA binds to DNA through its WD domain while contacting the ATPase domain of the neighbouring DnaA forming an higher order oligomer [Takeda and Dutta, 2005]. Similar observations, from the crystal structure of the heterodimeric *Sso*Orc1/Cdc6-1–3•DNA, suggested a similar higher order organisation [Dueber et al., 2007].

Origin licensing is achieved in two steps: recruitment and loading of the replicative helicase at the origin and the principal function of ORC, Cdc6 and Cdt1 is to carry out, in an ATP–dependent manner, the loading of the MCM2-7 complex (herein MCM2-7), hence forming the pre–RC [Bell and Dutta, 2002]. Several studies indicate that MCM2-7 is the replicative helicase during S phase [Labib et al., 2001;

Forsburg, 2004; Moyer et al., 2006].

Several studies in the past few years have revealed the events occurring during MCM2-7 loading. *In vitro* reconstruction experiments (loading assay), with purified ORC, Cdc6, Cdt1 and MCM2-7 revealed the events occurring during origin licensing [Frigola et al., 2013; Fernández-Cid et al., 2013; Randell et al., 2006; Remus et al., 2009; Takara and Bell, 2011] and a model of MCM2-7 loaded onto DNA as a double hexamer has been put forward [Samson and Bell, 2013].

Previous MCM2-7 loading assay with OCR, Cdc6, Cdt1, MCM2-7, showed that in the presence of ATP, MCM2-7 was loaded onto DNA in a high–salt stable complex. By contrast, in the presence of ATP$\gamma$S, no MCM2-7 was loaded onto DNA [Randell et al., 2006; Remus et al., 2009]. This suggested a two step loading, in which MCM2-7 is first recruited (high–salt unstable complex) and then loaded, in a ATP–dependent manner (high–salt stable complex) [Randell et al., 2006; Remus et al., 2009]. ORC associates with the origins in an ATP–dependent manner and in turn is needed for recruiting Cdc6. Additionally, ORC controls ATP binding and hydrolysis of Cdc6. ORC–Cdc6•DNA is required for recruiting Cdt1 and MCM2-7 to the origin via the interaction of Cdt1 with Cdc6 [Randell et al., 2006; Remus et al., 2009].

Recently, a Cdt1–independent MCM2-7 recruitment via ORC–Cdc6•DNA has been observed in the presence of ATP$\gamma$S. However, MCM2–4–6, the catalytic sub–complex of MCM2-7 [Ishimi, 1997; Kaplan et al., 2003; Lee and Hurwitz, 2001], was at substoichiometric levels [Frigola et al., 2013], hence suggesting that Cdt1 stabilises MCM2-7 during recruitment. Recruitment of the Cdt1–MCM2-7 complex [Remus et al., 2009] via the interaction with Cdc6 triggers ATP hydrolysis by Cdc6, releasing Cdt1 [Frigola et al., 2013; Randell et al., 2006]. Recruitment occurs via the interaction of the ORC–Cdc6•DNA bound with the C–terminal domain of Mcm3, which in turn stimulates ATP hydrolysis of ORC–Cdc6 [Frigola et al., 2013]. Interestingly, the C–terminus of Mcm3 is needed for loading both MCM2-7 hexamers [Frigola et al., 2013]. Hence an ORC–Cdc6–MCM2-7 complex could be an important intermediate in recruiting a second MCM2-7. Fernández-Cid et al. [2013] have shown that the C–terminal domain of Mcm6 has an autoinhibitory effect on the association of MCM2-7 with ORC–Cdc6. The interaction of Cdt1 with Mcmc6

[Yanagi et al., 2002; Liu et al., 2012] reduces this inhibitory activity, hence promoting MCM2-7 loading. ATP–hydrolysis by Cdc6 releases Cdt1 [Fernández-Cid et al., 2013]. Samson and Bell [2013] proposed a model by which MCM2-7 is loaded as double hexamer (Figure 1.3).

Similar *in vitro* loading assays with *Pfu*MCM and *Pfu*Orc1/Cdc6 suggested that another factor may be required for MCM loading since no high–salt stable complex was observed. Hence suggesting that only recruitment was occurring [Akita et al., 2010].

### 1.4.2 Origin activation

Origin activation is operated by two kinases, a cyclin dependent kinase (CDK) and a Dbf dependent kinase (DDK), which activate the pre–RC into pre–IC with subsequently recruitment of additional replication factors (Figure 1.2) [Bell and Dutta, 2002; Takeda and Dutta, 2005]. Interestingly, genome–wide transcription mapping indicates that serine–threonine protein kinases show cyclic induction in *Sulfolobus* species, suggesting that regulatory factors similar to eukaryotic CDKs may be present in archaea [Wu et al., 2014].

Additional recruited factors include the GINS complex and the Cell division cycle 45 protein (Cdc45). These factors are spatially associated with the origin and temporally regulated throughout the S phase [Takeda and Dutta, 2005].

GINS (Go–ichi–ni–san, 5–1–2–3 in Japanese) was originally discovered as an essential initiator factor in *S. cerevisiae* [Takayama et al., 2003]. GINS plays a key role in the normal progression of the replication fork during S phase [Gambus et al., 2006]. In this respect, GINS allows association of Cdc45 with MCM2-7 at the replicaiton fork [Gambus et al., 2006]. Cdc45, GINS and MCM2-7 form the 'CMG' complex (Cdc45–GINS–MCM2-7) or 'unwindosome' which localizes at the replication fork during DNA replication [Pacek et al., 2006; Moyer et al., 2006]. EM studies with *D. melanogaster* (*Dme*) CMG complex suggest that GINS and Cdc45 seal the gap between Mcm2 and –5, and that in the presence of ATP, the planar configuration of MCM2–7 is stabilised [Costa et al., 2011]. This induced conformational change would allow the appropriate juxtaposition of two functional

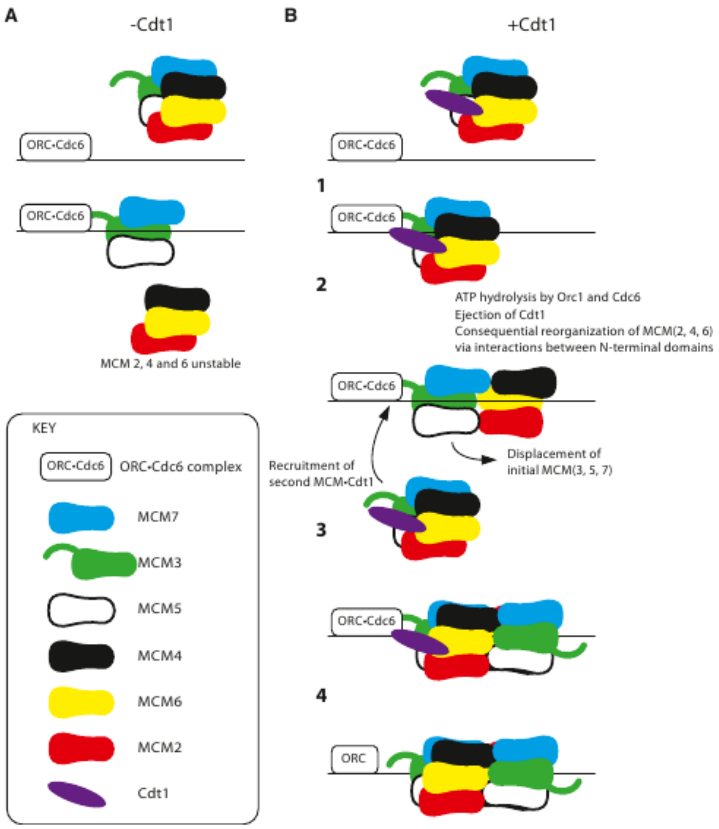



**Figure 1.3: Model for pre–RC formation.** (A) Cdt1–independent recruitment of MCM2–7 through the interaction of MCM3 C–terminus [Frigola et al., 2013]. (B) In the presence of Cdt1, the autoinhibitory effect of MCM6 C–terminus is reduced and the Cdt1–MCM2–7 can establish interactions with Orc–Cdc6•DNA bound [Fernández-Cid et al., 2013]. Upon ATP hydrolysis, Cdt1 is released with subsequent rearrangement of the MCM3/5/7 into an 'open–book' configuration. Another round of MCM2–7 recruitment and loading occurs in the same way, resulting in the formation of a double hexamer [Samson and Bell, 2013]. [Figure from Samson and Bell, 2013]

motifs (Walker A and Arg–finger) in both Mcm2 and 5 subunits and form the active MCM2-7 [Costa et al., 2011].

Electron microscope studies on the GINS complex showed a ring–like shape and, because of its stimulating activity on DNA pol–$\alpha$, it was suggested to function as a sliding clamp [Kubota et al., 2003; De Falco et al., 2006]. However, structural studies with the human heterotetrameric GINS complex revealed a tapered flat central cleft [Kamada et al., 2007]. Work in yeast, shed light on another factor, chromosome transmission fidelity protein 4 (Ctf4), which interacts directly with GINS and DNA pol–$\alpha$ coupling their activities [Gambus et al., 2009]. This interaction has been structurally determined and Ctf4 resolves to work as a bridge to couple the activity of GINS and DNA pol–$\alpha$ [Simon et al., 2014].

The archaeal GINS complex was originally identified in *S. solfataricus* in an attempt to identify interacting partners of *Sso*MCM [Marinsek et al., 2006]. *In vivo* and *in vitro* studies suggests that *Sso*GINS complex is a dimer of dimers consisting of two GINS proteins, GINS23 and GINS51, with GINS15 interacting with itself and GINS23 [Marinsek et al., 2006]. In *P. furiosus*, *Pfu*GINS complex stimulates *Pfu*MCM activity, which is otherwise very weak [Yoshimochi et al., 2008]. Similar results were reported for *Dme*MCM2–7, which in the absence of GINS complex and Cdc45 has weak helicase activity [Ilves et al., 2010]. It is possible that similar mechanisms of activation occur in archaea. The three dimensional structure of the archeal GINS complex from *T. kodakarensis* (*Tko*) revealed that the overall assembly of *Tko*GINS complex is a dimer of dimers similar to the human GINS, hence suggesting a similar role at the replication fork [Oyama et al., 2011].

Cdc45 was originally isolated as cold–sensitive mutant in yeast and it was shown to be essential for cell viability [Zou et al., 1997]. Previous sequence analysis suggested sequence similarities between the conserved DHH domain of Cdc45 and bacterial RecJ, which is a 5'–3' single–strand DNA exonuclease [Sanchez-Pulido and Ponting, 2011]. Homologs of RecJ have been identified in archaea [Marinsek et al., 2006], and it was suggested that archeal RecJ might have a similar role of Cdc45 at the replication fork.

Structural studies of a truncated form of *T. thermophilus* RecJ protein revealed a

typical DHH fold [Yamagata et al., 2002]. Biochemical characterisation of an active ssDNA DNA–specific 5'–3' exonuclease from *T. kodakarensis*, named GAN nuclease (GINS–associated nuclease) revealed an interaction with *Tko*GINS15, which in turn stimulates *Tko*RecJ's nuclease activity [Li et al., 2011].

To date, no functional activity has been reported for Cdc45. However recent data suggest a model whereby Cdc45 may act as a molecular wedge and help unwinding DNA duplex during DNA replication [Szambowska et al., 2014].

The initiation phase terminates with origin unwinding and recruitment of single–strand binding proteins, primases, loading of sliding clamps and DNA polymerases for initiating DNA synthesis (discussed below)[Takeda and Dutta, 2005].

### 1.4.3 The archaeal Mini–chromosome maintenance

MCM genes encoding proteins were first described in yeast as mutants whose mutations abolished the ability of the cells to maintain a plasmid containing a centromer and a replication origin [Maine et al., 1984; Takahashi et al., 1994].

MCM proteins are representative of a group of conserved nuclear proteins implicated in DNA replication of archaeal and eukaryal genomes [Kearsey and Labib, 1998]. In eukaryotes, the best known MCM proteins, are a family of six conserved proteins, which form the replicative helicase MCM2-7, directly implicated in the initiation and elongation step of DNA replication [Labib et al., 2001; Labib, 2000]. Although a few eukaryal MCM proteins have been identified to date, not all MCM proteins are directly involved in DNA replication [Maiorano et al., 2006; Kearsey and Labib, 1998].

Homologues of eukaryal MCM proteins have been identified in all archaeal genomes sequenced to date [Barry and Bell, 2006]. Conversely to eukaryotes, archaea have only one MCM gene homologous of the MCM2-7 family [Barry and Bell, 2006]. It has been proposed that MCM is also the replicative helicase in archaea [Sakakibara et al., 2009b; Barry and Bell, 2006].

*M. thermoautrophicum* was the first archaeal MCM protein whose properties were studied. Recombinant *Mth*MCM was shown to bind ssDNA and dsDNA; hydrolyse ATP, in the presence of DNA; and possess, 3'→5' ATP–dependent DNA

helicase [Kelman et al., 1999].

All archeal MCM proteins possess helicase activity and both DNA–dependent and –independent ATPase activity [Kelman et al., 1999; McGeoch et al., 2005; Barry et al., 2007; Atanassova and Grainge, 2008; Jenkinson and Chong, 2006; Fletcher et al., 2003] with the exception of *Mka*MCM, which is inactive [Bae et al., 2009]. Like the eukaryotic MCM4/6/7, archaeal MCM proteins have a 3'→5' helicase activity [Barry and Bell, 2006]. By contrast, the bacterial DnaB possesses a 5'→3' helicase activity [Kornberg and Baker, 1992].

EM studies suggest that archaeal MCM proteins adopt a double hexameric assembly, although hexameric single rings, heptameric single rings, double heptameric rings and filamentous form have been reported [Gómez-Llorente et al., 2005; Chen et al., 2005; Bae et al., 2009; Costa et al., 2006a; Pape et al., 2003; Slaymaker et al., 2013].

In archaea, the MCM proteins contain about 650 amino acids and can be divided in three structurally different domains: N–terminal, $AAA^+$ and helix–turn–helix (HTH) (Figure 1.4A).

### The architecture of the N–terminal domain of archaeal MCM proteins

The N–terminal domain is required for DNA binding and multimerisation of the MCM complex [Fletcher et al., 2003; Kasiviswanathan et al., 2004].

The N-terminal has also a role in regulation of the ATPase activity, helicase activity and substrate specificity. MCM proteins with deleted N–terminus have higher ATPase activity, robust helicase activity and altered range of substrate that can melt [Barry et al., 2007; McGeoch et al., 2005; Jenkinson and Chong, 2006].

The crystal structure of the N–terminal domain of *Mth*MCM and *Sso*MCM shows a three domain (A, B and C) architecture (Figure 1.4 A) and a central positively charged channel, large enough (23 Å) to accommodate either ssDNA or dsDNA [Fletcher et al., 2003; Liu et al., 2008]. The central channel is decorated with $\beta$–hairpins (C domain), which have been shown to bind both ssDNA and dsDNA [Fletcher et al., 2003; Kasiviswanathan et al., 2004; Poplawski et al., 2001; McGeoch et al., 2005]. The A domain plays a role in regulating helicase activity [Ka-

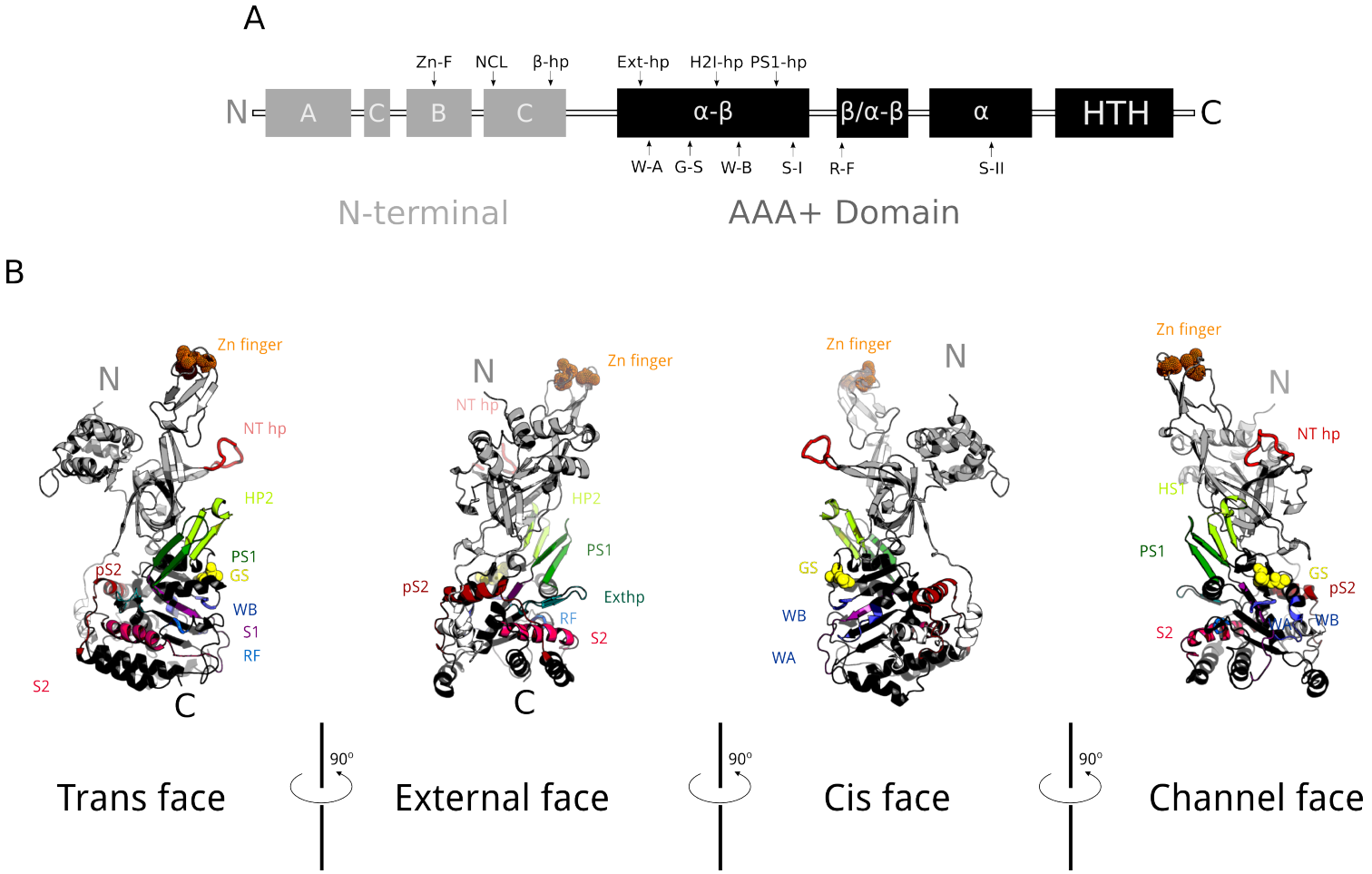



**Figure 1.4: Domain organization of the monomeric *Sso*MCM helicase.** (A) Cartoon showing the domain organization and the main structural motifs *Sso*MCM. The N–terminal part is divided into three domains, A, B and C. The N–terminal domain is linked to the AAA$^+$ catalytic part by an N–C linker. The AAA$^+$ part is divided into two domains, $\alpha/\beta$ and $\alpha$, which are connected by $\alpha/\alpha$–$\beta$ linker, which in turn is connected to the C–terminus helix–turn–helix domain (HTH). Arrows indicate the main structural motifs. Zn–F, zinc finger; NCL, N–terminal communication loop; $\beta$–hp, N–terminal $\beta$–hairpin; Ext–hp, $\beta$–hairpin on the exterior of the helicase; W–A, Walker–A; G–S, glutamate switch; H2I–hp, helix–2 insertion $\beta$–hairpin; W–B, Walker–B; PS1–hp, pre–sensor 1 $\beta$–hairpin; S–I, sensor–1; R–F, arginine finger; S–II, sensor–2. (B) Structures of the *Sso*MCM protein [Brewster et al., 2008].

siviswanathan et al., 2004] and recent studies provided evidence for a DNA binding site [Costa et al., 2008]. The B domain contains a zinc–finger motif ($C_4$ type), which binds to ssDNA. Mutations of this motif affect DNA–dependent ATP hydrolysis and helicase activity, but not the oligomeric structure [Kasiviswanathan et al., 2004; Poplawski et al., 2001]. The C domain is involved in multimerisation of MCM proteins into a ring [Kasiviswanathan et al., 2004]. Additionally, the C domain contains an highly conserved N–terminal communication loop (NCL) [Sakakibara et al., 2008]. Mutational analyses and structural studies demonstrated that the NCL has a key role in communication between N–terminal domain and the $AAA^+$ module of adjacent subunits within the MCM ring [Sakakibara et al., 2008; Barry et al., 2009].

### The architecture of the MCM $AAA^+$ catalytic domain

MCM proteins are member of the $AAA^+$ superfamily of ATPases. Members of this superfamily use the energy from cycles of ATP binding, hydrolysis and release of ADP to effect DNA melting and translocation along the DNA duplex [Ogura and Wilkinson, 2001a; Snider et al., 2008]. An important advance in understanding the structural biology of MCM proteins came with the crystal structure of the monomeric *Sso*MCM (residues 7–601, lacking the 85 HTH domain) [Brewster et al., 2008] (Figure 1.4 B) and, more recently, the structure of the monomeric full–length *Mka*MCM [Bae et al., 2009]. These structures revealed the $AAA^+$ catalytic domain folds into two distinct domains: an $\alpha/\beta$–domain and a three–strands helical domain called the $\alpha$–domain or lid domain [Brewster et al., 2008; Bae et al., 2009]. The two domains are connected by the $\alpha/\beta$–$\alpha$ linker (Figure 1.4 A) [Brewster et al., 2008; Bae et al., 2009]. Common structural motifs of the active site of $AAA^+$ superfamily of ATPases are: Walker A and B, sensor 1 and 2, and arginine–fingers, which carry out cycles of ATP binding, hydrolysis and release of ADP. The energy released for this process drives conformational changes in the motor that are required for the power stroke [Ogura and Wilkinson, 2001a; Snider et al., 2008].

Modelling of the *Sso*MCM crystal structure into the 3D–EM map of the double hexameric ring of *Mth*MCM, allowed the generation of the first models of a hexameric assembly of *Sso*MCM (Figure 1.5 A). This model showed that individual

subunits have a 'cis' and a 'trans' face (Figure 1.5 B). In this respect, the 'cis'–acting structural motifs are the Walker A, B and sensor 1 whilst sensor 2 and arginine–finger are the 'trans'–acting structural motifs [Brewster et al., 2008; Bae et al., 2009; Moreau et al., 2007].

The structure also revealed three $\beta$–hairpins: the H2I–hairpin, the PS1–hairpin, both pointing toward the central channel, while the third one, the Ext-hairpin is located exteriorly on the surface of the hexameric ring [Brewster et al., 2008].

The glutamate switch is another motif found in the $AAA^+$ domains which 'senses' the presence or absence of ligand (DNA in the case of MCM) and induces repositioning of the key glutamate residue in the Walker B thus regulate ATPase activity upon ligand binding control [Zhang and Wigley, 2008; Beattie and Bell, 2011].

### The C–terminal HTH domain

The C–terminal HTH domain might have a regulative role, in particular, a negative effect on the helicase activity of MCM complex [Barry et al., 2007; Jenkinson and Chong, 2006]. Deletion of the HTH domain showed increased ATPase and helicase activity [Barry et al., 2007; Jenkinson and Chong, 2006].

The MCM complex is a 3'→5' helicase and, as such, it translocates along DNA in a 3'→5' direction whilst melting DNA [Bell and Botchan, 2013]. The motor that drives the DNA melting process is the $AAA^+$ module. Cycles of ATP binding, hydrolysis and release within the active site of the $AAA^+$ module drive the conformational changes, within the MCM complex, required for effecting DNA unwinding [Bell and Botchan, 2013]. As, in other example of $AAA^+$ proteins, the active site in the MCM complex is located between the subunits forming the ring [Ogura and Wilkinson, 2001a; Snider et al., 2008]. This bipartite nature of the active site provides a mechanism for communication between subunits. Studies on *Sso*MCM have shown cooperation between subunits [Moreau et al., 2007]. Barry et al. [2009] suggested that this cooperation depends on the NCL. In particular, during cycles of ATP binding, hydrolysis and release of ADP, the PS1–hairpin, within the subunit, is repositioned to establish contact with the NCL in the neighbour subunits,

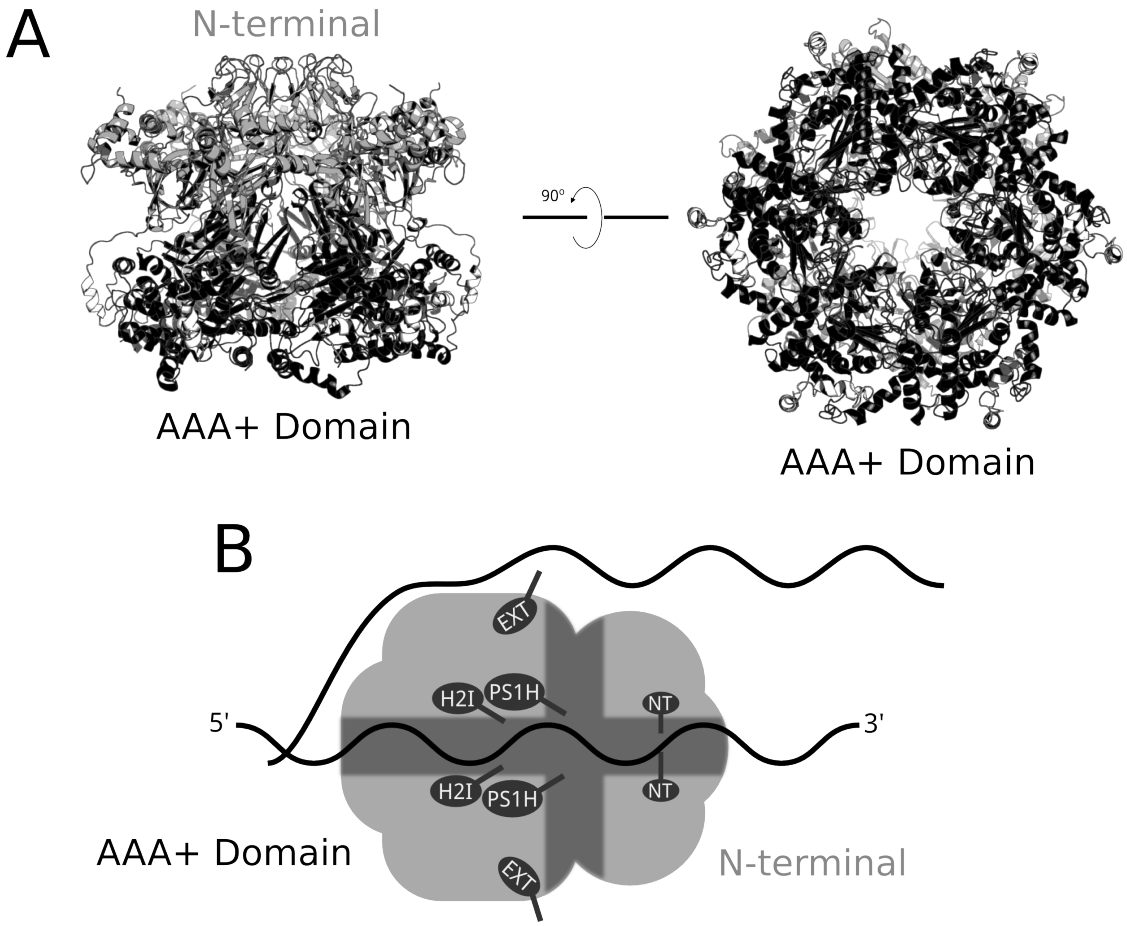



**Figure 1.5: Model of hexameric assembly of *Sso*MCM and proposed unwinding mode.** (A) Model of hexameric assembly of *Sso*MCM. Left, side view in which is visible the side channel between subunits. Right, the same hexamer rotated 90°to a top view down the central channel, looking from the C–terminal face. (B) Steric–exclusion model of a single *Sso*MCM helicase. DNA is shown as black lines. *Sso*MCM helicase is shown while translocating along the DNA with one single strand of DNA passing through the inner channel. The other strand is displaced ahead of the C–terminal domain [Bell and Botchan, 2013; Brewster et al., 2008]. Ext, $\beta$–hairpin on the exterior of the helicase; H2I, helix–2 insertion $\beta$–hairpin; PS1, pre–sensor 1 $\beta$–hairpin; NT, N-terminal $\beta$–hairpin;.

which in turn repositions its N–terminal $\beta$–hairpin. Evidence suggests that a second $\beta$–hairpin in the $AAA^+$, the H2I–hairpin, undergoes to nucleotide–depending repositioning [Jenkinson and Chong, 2006]. The crystal structure of *Sso*MCM showed that these to $\beta$–hairpins are in close apposition to each other, suggesting that these two elements may function as single module [Beattie and Bell, 2011].

To explain the helicase activity of the archaeal MCM complex, several models have been proposed [Sakakibara et al., 2009b]. However, data suggest a simple steric–exclusion model whereby a single hexameric MCM moves along ssDNA while displacing a second ssDNA ahead of it [McGeoch et al., 2005; Rothenberg et al., 2007; Graham et al., 2011; Bell and Botchan, 2013] (Figure 1.5 B).

## 1.5 DNA replication: elongation phase

The DNA duplex is replicated during the elongation phase, after two replisomes per origin have been assembled [Kornberg and Baker, 1992]. The antiparallel structure of DNA requires lagging strand synthesis to proceed in the opposite direction of the replication fork. Thus, during chromosomal DNA replication the leading strand is replicated continuously while the lagging strand is extended discontinuously in the opposite direction, respect to the leading strand, by short RNA–DNA fragments, the Okazaki fragment (Figure 1.6) [Kornberg and Baker, 1992].

At the onset of the elongation phase, the duplex DNA is opened and unwinded and the newly exposed ssDNA is coated with single–strand binding proteins (SSB). The DNA primase, which synthesizes the short RNA primers for the initiation of DNA synthesis, recruited to the replication bubble along with the replicative polymerases to initiate rapid and processive bidirectional DNA synthesis [Kornberg and Baker, 1992]. A crucial event in this phase is the polymerases switch from the primase to the high processivity polymerases, which, at least in eukaryotes, is most likely catalysed by the replication factor C (RFC) [Hubscher and Seo, 2001]. RFC is a DNA–dependent ATPase, which catalyses the opening of PCNA in order to load it on DNA duplex at the replication fork [Kelch et al., 2011] ([RFC are reviewed in Yao and O'Donnell, 2012]). It is known that PCNA confers processivity to replicative

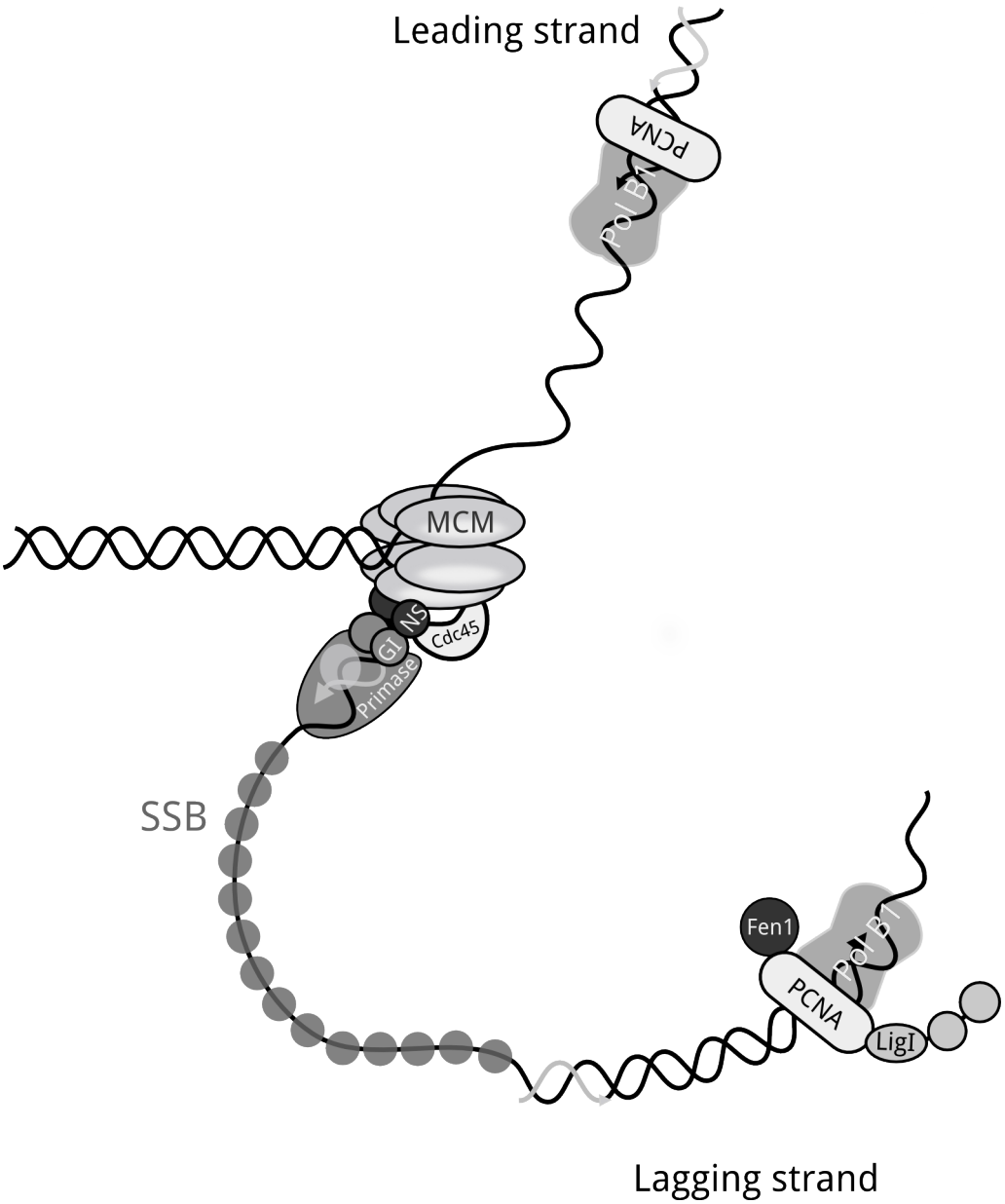



**Figure 1.6: Model of the architecture of replication fork in archaea.** DNA is indicated in black. Names of the enzymes are shown. RNA primers (in grey) are synthesised by primase. MCM is shown as a hexameric assembly with the leading strand exiting from the side channel of the ring. MCM intercats with GINS, which in turn interacts with the primase. Single–strand DNA binding proteins, binds the ssDNA a stabilise it during replication. DNA polymerases extend the RNA primers on both leading and lagging strand. PCNA can act as molecular 'tool belt' and 'carries' enzymatic activities (PolB1, Fen1 and LigI) for Okazaki fragment maturation along the lagging strand (Adapted from [Barry and Bell, 2006]).

polymerase [Kornberg and Baker, 1992].

In bacteria, the organisation of the replication fork has been well characterised [reviewed in Johnson and O'Donnell, 2005]. It is known that the DnaG, which associate with the replicative helicase DnaB, effects primase activity. Pol III is the high–processivity replicative polymerase and associate with the $\beta$–clamp [Kornberg and Baker, 1992; Johnson and O'Donnell, 2005]. Coordination of the replication fork occurs through the action of the $\gamma$–$\theta$ protein complex, which tethers and coordinates the enzymatic activities of DnaB, DnaG and Pol III, by physically interacting with them [Johnson and O'Donnell, 2005].

In eukaryotes, three polymerases are loaded: DNA pol–$\alpha$ (primase), DNA pol–$\delta$ and DNA pol–$\epsilon$, although Pol–$\delta$ can accomplish the synthesis of both lagging and leading strand without Pol$\epsilon$ [Takeda and Dutta, 2005; Hubscher and Seo, 2001]. Additional factors needed for efficient DNA synthesis include topoisomerases, which relieve tension created by the replication fork in the duplex DNA. Additionally, a number of enzymes are required for maturation of the Okazaki fragments to form a continuous duplex DNA [Kornberg and Baker, 1992].

In terms of proteins participating in the elongation of the lagging and leading strand, the archaeal machinery is a simplified version of that of eukaryotes (Figure 1.6) [Barry and Bell, 2006; Beattie and Bell, 2011; Ishino and Ishino, 2012; Lindås and Bernander, 2013]. In archaea, the homologue of DNA pol–$\alpha$ is the heterodimeric primase [Barry and Bell, 2006]. In eukarya, the primase is a heterotetramer composed of PriS, PriL, Pol–$\alpha$ and B subunit [Johnson and O'Donnell, 2005]. The archaeal primase is a heterodimer composed of a small (PriS) and a large (PriL) subunit homolog of eukaryal PriS and PriL [Barry and Bell, 2006]. In *S. solfataricus*, *P. furiosus* and *P. horikoshii* PriS possesses both RNA and DNA polymerase activity *in vitro* [Barry and Bell, 2006]. PriL seems to have a regulative role with respect to PriS since it was observed that in the presence of PriL, PriS had increased RNA polymerase activity and a more homogeneous average product length [Barry and Bell, 2006]. Archaeal genomes encode DNA dependent DNA polymerases (discussed in section 1.5.3), however little is known about archaeal DNA polymerase *in vivo* [Barry and Bell, 2006]. As in eukaryotes, PCNA is loaded, at the replication

fork, in an ATP–dependent manner by RFC [Lindås and Bernander, 2013]. Similarly to eukaryotic polymerases, archaeal polymerases associate with PCNA via the PIP–box and association results in increased polymerase processivity [Beattie and Bell, 2012].

### 1.5.1 Okazaki fragment maturation

Okazaki fragment maturation represents a required procedure to impart chemical stability to the lagging strand [Kornberg and Baker, 1992]. Due to the antiparallel configuration of the DNA duplex, the synthesis of the lagging strand involves some elegant enzymatic choreography [Nelson et al., 2008]. While the leading strand is replicated continuously, the lagging strand is replicated discontinuously by short RNA–primed fragments termed Okazaki fragments [Kornberg and Baker, 1992]. Okazaki fragments are then joined together in a process called Okazaki fragment maturation. Okazaki fragments are generated through strand displacement by the polymerase. In this process structure specific flap endonuclease 1 (Fen1) recognises the DNA flap and cleaves it and thus leaving a nicked duplex DNA. Ligase I recognises this nick and seal it off [Kornberg and Baker, 1992].

Archaeal and eukaryal Okazaki fragments are similar in length (∼100 nt) [Matsunaga et al., 2003], whereas bacterial Okazaki fragments are longer (∼ 1000 nt) [Kornberg and Baker, 1992]. The crenarchaeon *S. solfataricus* has been adopted as model organism for Okazaki fragment maturation since it was shown to possess a simplified tool set of eukaryotic–like Okazaki fragment maturation proteins (DNA polymerase, Fen1, LigI and PCNA) [Dionne et al., 2003; Beattie and Bell, 2012].

*In vitro* Okazaki fragment maturation experiments, showed that *Sso*PCNA, *Sso*PolB1, *Sso*Fen1 and *Sso*LigI are necessary and sufficient for an efficient for Okazaki fragment maturation [Beattie and Bell, 2012]. Interestingly, previous *in vitro* experiments with the same proteins showed that a single *Sso*PCNA ring could bridge, between *Sso*Fen1 and *Sso*Lig1 or *Sso*PolB1, and binding was specific for each one. In particular, PolB1 binds to PCNA1; Fen1 binds to PCNA2; whereas Lig1 binds to PCNA3 [Dionne et al., 2003]. These data led to hypothesise a model in which PCNA establishes multiple simultaneous interactions with its interactors,

hence forming a molecular 'tool belt' [Dionne et al., 2003].

### 1.5.2 The proliferative cell–nuclear antigen

Proliferative cell–nuclear antigen (PCNA) was historically identified as a nuclear antigen, highly expressed during S phase in proliferating cells [Warbrick, 2000]. PCNA proteins also participate in DNA repair and recombination [reviewed in Moldovan et al., 2007]. PCNA belongs to the family of DNA sliding clamps ($\beta$–clamp) [Moldovan et al., 2007]. Sliding clamps are ring–shaped proteins whose central cavity is large enough to accommodate dsDNA hence encircle DNA and allow tethering of their cognate client proteins to the DNA template (Figure 1.7) [Jeruzalmi et al., 2002]. As an example, tethering the polymerase by sliding clamps enables processivity and high-speed replication during chromosomal DNA replication. Sliding clamp proteins are structurally and functionally conserved through the three domains of life [Moldovan et al., 2007; Warbrick, 2000; Jeruzalmi et al., 2002]. PCNA sliding clamps have no known activity and the only activity known is its ability to move along dsDNA [Pan et al., 2011].

Eukaryotic PCNA is a homotrimeric complex while the bacteria homologue is a homodimeric complex termed $\beta$–clamp [Moldovan et al., 2007]. The crystal structure of human PCNA [Krishna et al., 1994] revealed the ring–shaped trimeric complex with three PCNA monomers arranged in a head–to–tail manner. PCNA monomers are divided into two structurally similar domains (domain I and II), which are connected via a long stretch of residues termed the inter–domain connecting loop (IDCL) [Mailand et al., 2013]. In yeast PCNA•DNA bound structure, PCNA encircles dsDNA and can track freely along it in both directions [McNally et al., 2010].

Archaeal PCNA forms homotrimeric rings in Euryarcheaota (i.e. *P. furiosus*), whereas in Crenarchaeota, rings are heterotrimeric (i.e. *S. solfataricus*) (Figure 1.7) [Pan et al., 2011]. A number of archaeal PCNA proteins have been structurally characterised and revealed a striking structural similarity to the eukaryotic PCNA [Krishna et al., 1994].

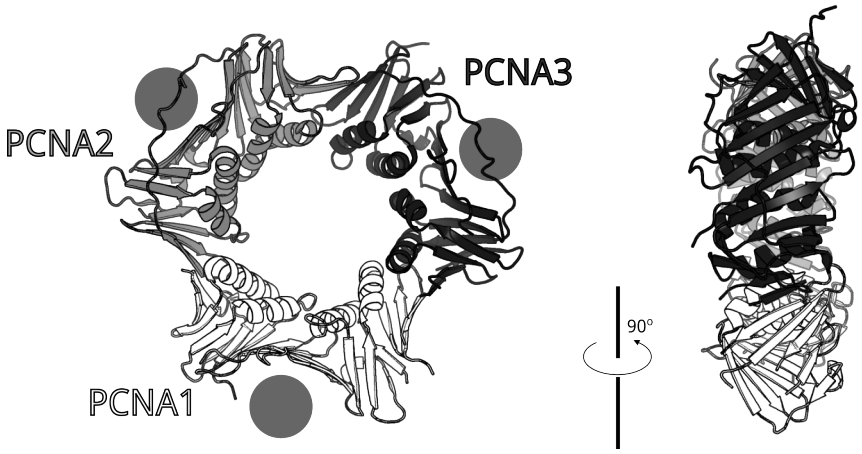



**Figure 1.7: Architecture of the *Sso* proliferative cell–nuclear antigen.** Left, Model of the heterotrimeric *Sso*PCNA. View down the central cavity of *Sso*PCNA. Monomers are indicated as PCNA1, PCNA2, PCNA3. Grey circles indicate the inter domain connector loop (IDCL), which is need to establish interactions with the PIP box of PCNA–interacting proteins. Right, the same model rotated of 90°.

### PCNA–interacting proteins, conformation flexibility and coordination of multiple proteins

The eukaryotic PCNA can establish several interactions with a large number of proteins (herein client proteins) and regulates their activities [summarised in Moldovan et al., 2007]. PCNA–interacting proteins bind to the IDCL, on PCNA (Figure 1.7), through the PCNA–interacting protein domain (PIP box) on the client protein [Moldovan et al., 2007; Pan et al., 2011]. The PIP box may be found either at the N–terminus or at the C–terminal tail of the client proteins [Moldovan et al., 2007]. The PIP box can be identified with the consensus sequence QxxΨxxΘ (Ψ, hydrophobic residues L, M or I; Θ, aromatic residues F or I) [Moldovan et al., 2007; Warbrick, 2000]. The first structure of the human PCNA in complex with the C–terminal tail of $p21^{WAF1/CIP1}$ showed that all the three subunits of the homotrimeric complex could establish extensive interactions with each of the subunits composing the ring [Gulbis et al., 1996].

Structural observations of human PCNA–$(Fen1)_3$ complex [Sakurai et al., 2005] and, more recently, the archaeal PCNA–$(RNase\ HII)_3$ complex [Bubeck et al., 2011] shown simultaneous binding without steric hindrance and three different conformations of identical client proteins [Sakurai et al., 2005; Bubeck et al., 2011]. A key structural motif, the 'hinge' motif, located between the PIP motif and the catalytic core of the client protein allowed such flexibility [Sakurai et al., 2005; Bubeck et al., 2011; Beattie and Bell, 2011]. Conformational flexibility has been observed in other structural studies of PCNA–client protein complexes [Mayanagi et al., 2011; Nishida et al., 2009; Pascal et al., 2006]. The DNA polymerase can adopt different conformational states upon a single PCNA ring, to allow switching between polymerising and editing mode [Mayanagi et al., 2011; Nishida et al., 2009]. Similar conformational flexibility was observed for LigI, which can adopt an extended conformation, in the absence of nicked dsDNA [Pascal et al., 2006], or a C–shaped conformation, in the presence of a nicked substrate [Nishida et al., 2009], upon a single PCNA ring. This observation suggested that client proteins are able to adopt flexibility around a single PCNA ring, which in turn facilitates the coordination of enzyme activity, acting as molecular 'tool belt' [Sakurai et al., 2005; Bubeck et al., 2011; Beattie and

Bell, 2011].

Multiple protein coordination is best supported by *Sso*PCNA [Beattie and Bell, 2011], which assembles in a stable homotetrameric ring [Williams et al., 2006] and has been shown to posses subunit specificity for distinct client proteins [Dionne et al., 2003]. All these observations raise the possibility that all three proteins can simultaneously associate with a single PCNA ring [Dionne et al., 2003; Beattie and Bell, 2011, 2012].

### 1.5.3 DNA Polymerase

DNA–dependent DNA polymerases carry out the synthesis of the new daughter DNA strand using RNA–primed ssDNA template and deoxynucleotides [Kornberg and Baker, 1992]. Polymerases are also involved in other processes within the DNA metabolisms, for example, repair and recombination. Based on sequence homology, DNA polymerases can be classified into six distinct groups (type (or family) A, B, C, D, X, and Y) [Hubscher et al., 2002; Joyce and Steitz, 1994].

In bacteria, two polymerases are entailed for DNA replication, Pol I, which belongs to the family A of polymerases and is involved in Okazaki fragment maturation; Pol III, which belongs to the family B of polymerases and carries out the synthesis of the leading strand [Kornberg and Baker, 1992].

In eukaryotes, the three polymerases essential for DNA replication belong to the family B polymerases. They share a conserved catalytic core although they have distinct function in S phase [Waga and Stillman, 1998; Takeda and Dutta, 2005; Hubscher and Seo, 2001].

Only the B and D family of polymerases have been found in archaeal genomes (Figure 1.8) [Barry and Bell, 2006]. Three groups of family B polymerases have been identified in all archaeal genomes [Grabowski and Kelman, 2003]. The family D polymerases seem to be unique to the euryarchaeota phylum [Cann et al., 1998]. *Pfu*PolD polymerases are composed of two subunits: DP1 and DP2. DP1, has sequence homology with the non–catalytic B subunit of the family B polymerases (Pol–$\alpha$ (p70 subunit), Pol–$\delta$ (Cdc27p) and Pol$\epsilon$ (p55 subunit)) [Grabowski and Kelman, 2003]. DP2, has distinct sequence from other DNA polymerases, contains the

polymerization activity [Grabowski and Kelman, 2003]. By contrast, in crenarchaea no D family of polymerases have been found [Barry and Bell, 2006].

It is not clear yet which DNA polymerase is the replicative polymerase in archaea. In euryarchaeota, biochemical data support a model where PolB synthesises the leading strand while PolD the lagging strand [Henneke et al., 2005]. In Crenarchaea, it may be possible that the three B–type of polymerase have distinct roles on the leading and lagging strand [Barry et al., 2007].

### 1.5.4 Flap-endonuclease 1

Flap-endonuclease 1 (Fen1) is a structure–specific endonuclease 5'–3' exonuclease whose activity is released on branched DNA molecules resulting in the cleavage of the 5' end single strand flap [Liu et al., 2004]. The DNA flap is a forked DNA structure composed of double-stranded DNA and a displaced single-strand (Figure 7.10A). A DNA flap is generated, during DNA replication of the lagging strand, following strand displacement of the downstream Okazaki fragment [Hubscher and Seo, 2001]. Fen1 plays a key role in Okazaki fragment maturation since it catalyses RNA primer removal [Turchi et al., 1994], in both eukarya and archaea [Hubscher and Seo, 2001; Grabowski and Kelman, 2003]. In bacteria, RNA primer removal it is carries out by Pol I [Kornberg and Baker, 1992].

Homologs of eukaryal Fen1 have been identified in archaeal genomes [Hosfield et al., 1998; Matsui, 1999; Dionne et al., 2003]. Sequence analysis of *M. jannaschii*, *A. fulgidus* and *P. furiosus* revealed a striking sequence similarity (75%) with the human Fen1 [Hosfield et al., 1998].

Biochemical and structural studies of archaeal homologues of Fen1 from several archaea shed light on biochemical properties, similar to the eukaryotic Fen1 in terms of substrate specificity and catalytic activity [Liu et al., 2004].

The three-dimensional structures of the *P. furiosus* [Hosfield et al., 1998], *M. jannaschii* [Hwang et al., 1998], and *P. horikoshii* [Matsui et al., 2002] Fen1 revealed a common feature, which is a long flexible loop large enough to accommodate ssDNA, supporting the biochemical observations of a threading mechanism for the substrate through the hole in the enzyme [Sayers and Artymiuk, 1998; Grabowski and Kelman,

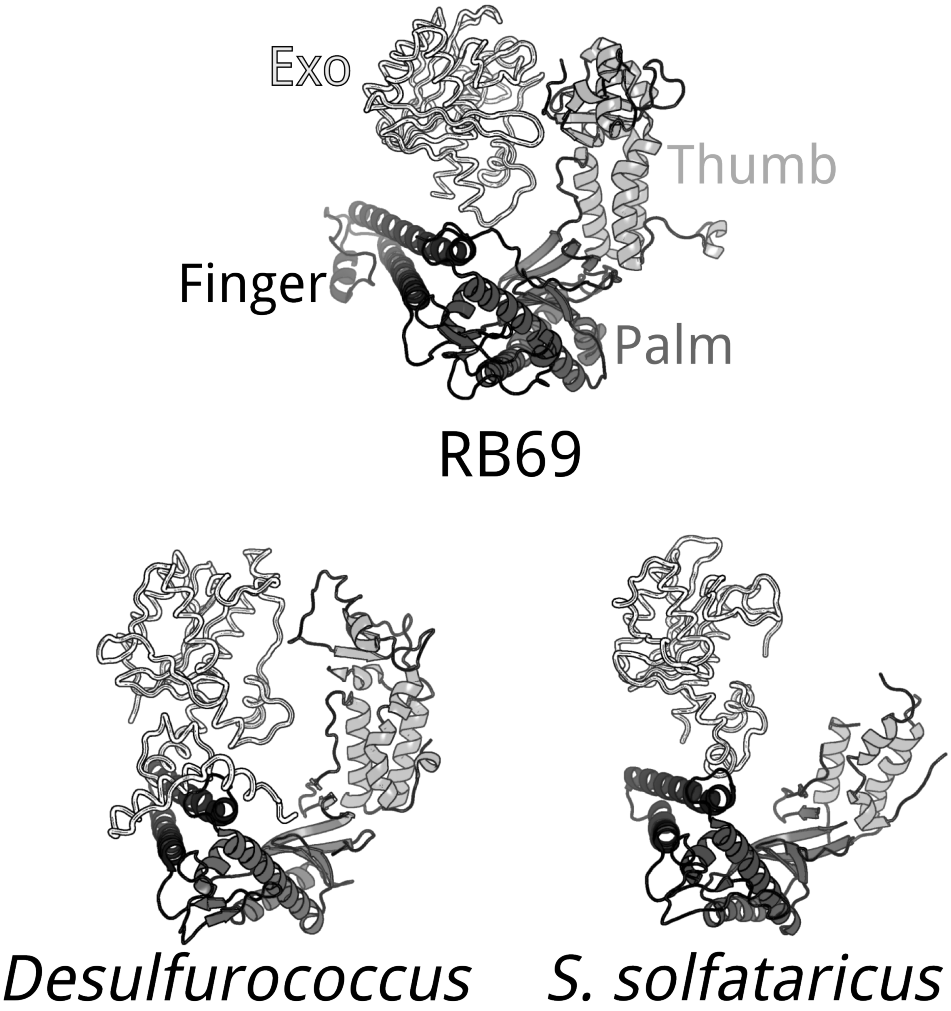



**Figure 1.8: Architecture of family B DNA polymerase.** Family B of DNA polymerases share a common overall architectural feature. They have a shape that can be compared with that of a right hand and have been described as consisting of 'thumb', 'palm' and 'fingers'. 'Palm' domain seems to be catalysis of the phosphoryl transfer reaction. 'Fingers' domain includes important interactions with either the incoming nucleoside triphosphate and the template base to which it is paired. 'Thumb' may play a role in positioning the duplex DNA and in processivity and translocation [Joyce and Steitz, 1994]. (*Desulfurococcus* strain Tok, PDB 1QQC; *S. solfataricus* PolB1, PDB 1S5J and apo RB69, PDB 1IH7).

2003].

### 1.5.5 DNA ligase I

A DNA ligase is an enzyme, which catalyses the formation of a phosphodiester bond between adjacent 3'–OH end and 5'–$PO_4$ termini of two DNA strands [Kornberg and Baker, 1992]. DNA ligases can be divided into two groups: ATP– and $NAD^+$–dependent ligases [Kornberg and Baker, 1992]. DNA ligases have key roles in many important cellular process including DNA replication, recombination, and repair [Kornberg and Baker, 1992]. DNA ligases, using ATP as a cofactor are mostly found in eukaryotes whereas, NAD–dependent ligases are normally found in bacteria [Shuman, 2009].

Homologues of the eukaryal ATP–dependent DNA ligase I (LigI) have been isolated in *M. thermautotrophicus* [Sriskanda et al., 2000], *Aeropyrum pernix* [Jeon and Ishikawa, 2003], *S. solfataricus* [Dionne et al., 2003], *P. furiosus* [Nishida et al., 2006]. Structural observation revealed that archaeal LigI can adopt either an extended (*Sso*LigI [Pascal et al., 2006]) or a C-shaped protein clamp conformation (*Pfu*LigI [Nishida et al., 2006]) (Figure 1.9). The latter, strongly resembles the structure of two well–characterised ligases, the human LigI [Pascal et al., 2004] and the bacterial LigA [Shuman, 2009], engaged onto nicked DNA (Figure 1.9).

## 1.6 Transmission electron microscopy and single particle analysis

In 1931, two Germans, the physicist Ernst August Friedrich Ruska, Nobel prize laureate (1986), and the electrical engineer Max Knoll built the first prototype of a transmission electron microscope (TEM), which was capable of magnifying an object up to 400 times. Since then, much has been achieved in terms of technology. Today, the state–of–art electron microscope can magnify an object to 50 million times. TEMs are so powerful, in terms of *resolution* or more accurately *resolving power*, since electrons, which have shorter wavelength than visible light, are used

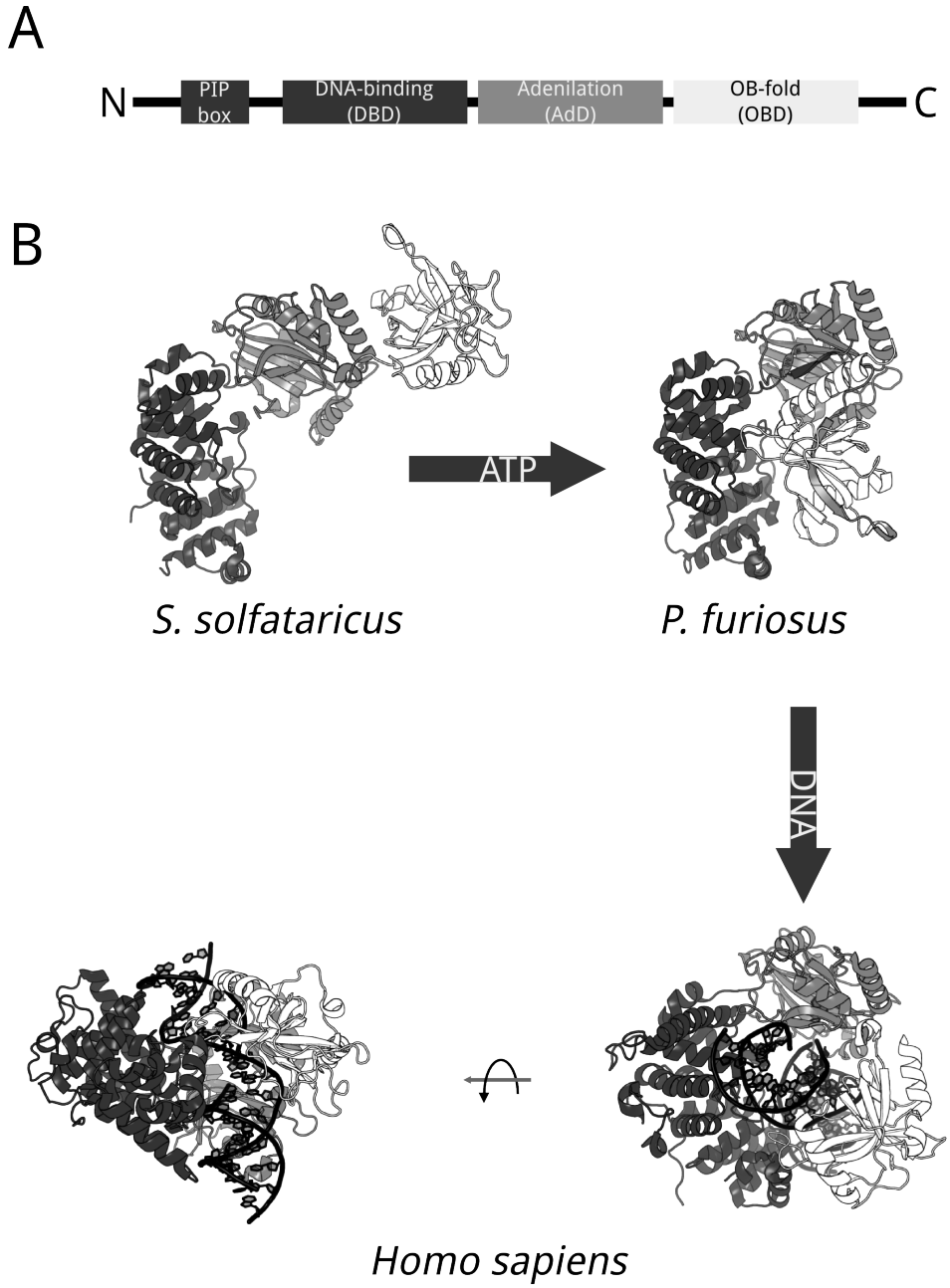



**Figure 1.9: Architecture and conformational changes of DNA ligase upon DNA binding.** (A) Linear representation of DNA ligase I domains. (B) Three crystal structures supporting DNA ligase I conformational changes. DNA ligases from *S. solfataricus* (PDB code 2HIV) adopt an extended conformation, *P. furiosus* (PDB code 2CFM) adopt a C–shaped conformation upon ATP binding similar to human LigI engaged onto DNA (PDB code 1X9N). The ATP–dependent DNA ligase I, consisting of three distinct domains, changes its conformation upon engaging a nicked DNA substrate and catalyse the ligation reaction.

as source of light to 'see' an object [Williams and Carter, 2009]. An easy way to express the *resolving power*, meant as '*the smallest distance that can be resolved*', of a TEM instrument is to use the Rayleigh criterion:

$$\delta = \frac{0.61\lambda}{\mu\ sin\theta} = \frac{0.61\lambda}{A}$$

Where

$\delta$: resolution limit

$\lambda$: electron wavelength (nm)

$\mu$: diffraction index, which is 1 in high–vacuum

$\theta$: semi–angle of collection of the magnification lenses

A: numeric aperture

The formula suggests that the resolving power or resolution limit ($\delta$) of a microscope is limited by the wavelength used to 'see' the object. Since electrons are smaller than atoms, it should be possible, theoretically, two 'see' below the atomic level [Williams and Carter, 2009]. However, the resolution limit of a TEM instrument is not limited by the wavelength but principally by the imperfections of the electron lenses (spherical and chromatic aberration). In fact, if microscopes were totally devoid of defects, their resolution would only be limited by the wavelength of the beam [Williams and Carter, 2009]. However, modern TEM instruments use hardware correction of lenses' imperfections and it is now possible to visualise an object at sub–Å scale [Williams and Carter, 2009].

Similar to a conventional light microscope, a TEM instrument consists of a light source, in which case is the electron source; lenses, which consist of electromagnetic rings that deflect electrons by an electromagnetic field; and an image detector, which can be a fluorescent viewing screen, a photographic film; a digital camera [Williams and Carter, 2009; Orlova and Saibil, 2011] (Figure 1.10 TEM). In this respect, electrons are produced by an electron gun, which can be either a conventional thermionic emitters (tungsten or $LaB_6$) or a field emission gun (FEG). The latter is better since produces an electron beam, which is smaller in diameter, more coherent and brighter than conventional thermionic emitters. Emitted electrons from the electron gun are accelerated by an anode. The resulting electron beam travels

through the microscope column and is delivered to the specimen and focused on the screen by electromagnetic lenses (Figure 1.10 TEM).

Electrons are ionizing radiation, which is the general term given to radiation capable of removing the tightly bound, inner-shell electrons from the attractive field of the nucleus [Williams and Carter, 2009]. Although a wide range of signals are generated during the interaction of the electron beam with the sample, the interaction of the electrons with the inner-shell electrons of the nucleus of the specimen results in scattering of the electrons [Williams and Carter, 2009; Orlova and Saibil, 2011]. More specifically, the collision of beam electrons with the electron cloud or nuclei of the specimen, leads to energy loss (inelastic scattering), while the deflection of the electron beam by the electron cloud, does not change the energy of the electron (elastic scattering) [Orlova and Saibil, 2011].

In TEM, an images is a representation of the intensity variations caused by local variations in the specimen transmission during the collision of the electron beam with the specimen [Orlova and Saibil, 2011]. The image is formed due to the transmitted scattered and unscattered electrons through the specimen [Orlova and Saibil, 2011]. The image contrast that results from the adsorption of part of the incident beam, and hence due to inelastic scattering, is known as amplitude contrast [Orlova and Saibil, 2011]. Biological samples adsorb only a small fraction of the electrons during inelastic scattering, however the amplitude contrast can be increased by using objective lens aperture to eliminate electrons scattered at high angles [Orlova and Saibil, 2011]. The image contrast that result from the interference of the elastically scattered electron with the unscattered electrons, is known as phase contrast [Orlova and Saibil, 2011]. The resulting image is visualised onto a fluorescent viewing screen or recorded onto a photographic film; a CCD (charge-coupled device) camera; or a direct detector (Figure 1.10 TEM).

### 1.6.1 TEM of biological sample

To fully understand the machinery of the living cells and their biological processes, it is essential to understand the structure and dynamics of the constituent molecules, how they assemble into complex molecular machines, and how they form functional

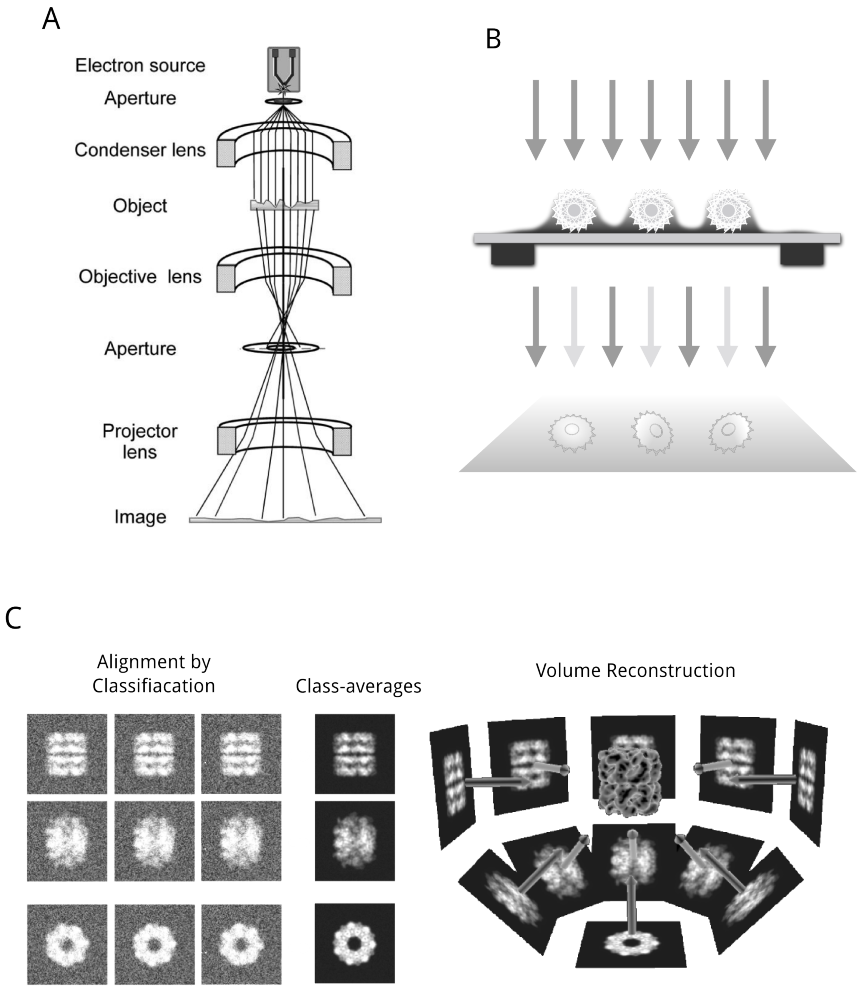



**Figure 1.10: Schematic representation of transmission electron microscopy and single particles analysis.** (A), Simplified schematic representation of an electron microscope [Figure from [Orlova and Saibil, 2011]]. (B), Negative staining principle. (C), Single particle analysis and 3D reconstruction example [Adapted from Dr Greg Pintilie's images at http://people.csail.mit.edu/gdp/cryoem.html].

organelles, cells, and tissues [Orlova and Saibil, 2011]. Since TEM's invention, EM has become an important tool in structural biology since by this approach it is possible to explore a broad range from atomic to cellular structures [Orlova and Saibil, 2011].

The first sub–nanometer cryo-EM single-particle reconstructions was determined for the icosahedral Hepatitis B virus capsid at 7.4 Å. For the first time, it was possible to recognize secondary structure elements as well as distinguish individual domains and trace the protein backbone [Böttcher et al., 1997]. Since then there have been an increasing number of structures determined by cryo-EM and single particle reconstruction in the range of 6–10 Å[Lindert et al., 2009; Orlova and Saibil, 2011].

Today, state–of–art TEM operating at 300–400 keV along with a new generation of electron detectors plus software for statistical analysis of molecular images, allows 3D reconstruction at atomic resolution of ∼3.2 Å [Kühlbrandt, 2014]. Routine limits for electron microscopy and single particle analysis are in the range of 7–30 Å. In this range of resolution it is possible to visualise secondary structure elements to the shape of the molecular assembly. In particular, $\alpha$–helices are visible as rods in the range of 7–10 Å whereas $\beta$–sheets and loops are difficult to discern [Lindert et al., 2009; Orlova and Saibil, 2011]. To visualise $\beta$–sheets and loops higher resolution (4–5Å) is needed [Lindert et al., 2009; Orlova and Saibil, 2011]. Lower resolutions, in the range of 10–15Å, give information about the domains organisation, whereas lower resolutions provides informations about the overall shape of the molecular complex.

### Electron microscopy of stained samples

Biological specimens can be barely seen under normal conditions of TEM, thus some treatments to enhance contrast are needed. For this purpose, heavy metal stains are used for increasing contrast in a micrograph. This heavy metals simply increase contrast owing to the fact that bigger atoms scatter electrons better [Hyatt, 1981]. When heavy metals are used to increase contrast, the image of the specimen on resulting photographic negative, appears dark against a light background. The term "negative staining", which is not staining at all since the specimen is embedded in

the staining solution, is "negative" only in the sense that the biological material has less contrast than the surrounding material. This method is called negative staining and it was first introduced by Brenner and Horne in 1959 [Frank, 1996]. This method has been extensively used to obtain molecular images of macromolecules with high contrast [Frank, 1996]. Stains mostly cause elastic scattering of electrons (Figure 1.10 NEGATIVE STAINING PRINCIPLE)[Hyatt, 1981]. The most commonly used heavy metal staining in TEM is the uranyl acetate. With uranyl stain, the possible resolution achievable under normal conditions is limited to about 15 Å. The highest resolution possible is obtained with negative stains such as uranyl formate, which has smaller size [Hyatt, 1981].

The actual chemical composition of the stain has little effect on contrast but the 'health' of the biological sample. High ion strength, extremely low pH as well as distortion of the shape due to sit–drying can alternate the molecular structure of the macromolecular complex being investigate through single particle electron microscopy. Nevertheless, single particles electron microscopy of samples embedded in stain is still the technique of choice to explore macromolecular protein complexes where X–ray crystallography an NMR are not applicable [Frank, 1996].

### Single particle analysis and 3D reconstruction

Single Particle Analysis (SPA) refers to image processing techniques used to analyse molecular images collected from TEM [van Heel et al., 2000]. Images of stained or unstained biological samples are very noisy and they are difficult if not impossible to interpret. In order to improve the quality of the image in terms of details that can be seen several approaches have been developed [van Heel et al., 1996; Frank et al., 1981; Scheres, 2012]. These approaches are based on statistical analysis and alignment technique of the images [van Heel et al., 2000]. The processing of the single–particle images aims to group together images with the same orientation and produce an averaged new image with improved S/N ratio, in order to determine translational and rotational parameters for each of the individual particles in the data sets. By averaging techniques, the S/N ratio is improved in the single image, leading to an image that can be interpreted and used for a 3D reconstruction.

Single molecules in solution are in random orientations, and they thus have six degrees of freedom: three translational ones (X, Y, Z) and three rotational ones, which correspond to three 'Euler angles' ($\alpha$, $\beta$ and $\gamma$). The image of the biological sample created in the electron microscope is a projection along the Z direction. These 2D images are essentially the 2D projections of the 3D electron potential of the specimen [Orlova and Saibil, 2011] (Figure 1.10). In order to calculate a 3D map from a set of 2D projections, the relative orientations must be determined [Orlova and Saibil, 2011]. There are different approaches used to assign relative orientations to a set of projections. The best approach is chosen based on the availability of an initial model. When the initial model as to be calculated, two general approaches are available.

In the first approach, called random conical tilt (RTC), involves recording of images, of the same particles, at different tilt angles. Particles pairs, corresponding to the same object at two different tilts are selected, and subsequently aligned and classified. The 3D reconstruction can be performed relatively easily since the Euler angles of each particle is known from the tilt geometry [Orlova and Saibil, 2011].

The second approach, called common lines, is computationally based and in this case only untilted images are collected. For this approach, it is necessary to collect a range of views distributed over different orientations [Orlova and Saibil, 2011]. This method is based on the observation that each pair of 2D projections of a given 3D structure have al least one 1D projection in common [Van Heel, 1987; Orlova and Saibil, 2011]. This approach can be implemented either in Fourier space and real space. The common line, between two 2D projection in Fourier space is the intersect of the two corresponding planes in Fourier space [Orlova and Saibil, 2011]; whereas the common line in real space is based on the concept of Radon transform. The Radon transform describes an N–dimensional function by a set of 1D projections. In the case of a 3D object this concept is particularly useful to reconstruct a 3D object from 2D projections. In particular, the Radon transform of a 3D object corresponds to the set of 1D projections of a section of the 3D object [Orlova and Saibil, 2011].

The third approach is called projection matching. If an initial model is available,

even at very low resolution, then the procedure of projection matching could be used. In this case the initial 3D model can be used to generate re–projections of all possible orientations. The set of re–projections can be used as reference images, in order to systematically compare each image in the data set with all of the reference images [Orlova and Saibil, 2011]. By this approach, the Euler angles of the reference image that gives the best cross–correlation are assigned to the raw image or class average. Once the Euler angles are assigned, a new 3D map can be calculated and the procedure iterated with the new set of re–projections [Orlova and Saibil, 2011]. This method is particularly useful for refinement of a 3D model, and hence improve resolution, once the initial model as been calculated.

## 1.7 Concluding remarks

Archaea are interesting organisms for studying molecular and evolutionary aspects of biology. Structural and biochemical studies have showed great similarity between eukaryal and archaeal replisomes[Ishino and Ishino, 2012]. Archaea possess multiple origins of replication, few initiator proteins (MCM and Cdc6/Orc1) and a simplified tool set of proteins (PCNA, DNA polymerase, Fen1 and ligase I) sufficient for Okazaki fragment maturation and hence represent a valid model for better understanding the eukaryal replisome [Barry and Bell, 2006]. Proteins from the hyperthermophilic archaea are more stable than their mesophilic homologues and hence more suitable for functional and structural analysis of multiple protein complexes [Ishino and Ishino, 2012].

Recent work in yeast revealed the molecular basis by which the replicative helicase MCM2-7 is recruited and loaded, as double hexamer, onto DNA at the origin of replication [Remus et al., 2009; Randell et al., 2006; Fernández-Cid et al., 2013; Frigola et al., 2013].

These mechanisms are still unknown in archaea. Data suggest a similar mechanisms may take place [Akita et al., 2010]. However, no archaeal Cdt1 has been reported yet even if a possible candidate may be Whip a novel initiator protein identified in *S. solfataricus* [Robinson and Bell, 2007].

In conclusion, the third domain of life offers numerous new opportunities for understanding the molecular events of chromosomal DNA replication.

# Chapter 2

# Aims of The Projects

Much of what we know about DNA replication comes from studies concerning the functional and structural characterization of proteins involved in DNA replication in bacteria, but less concerned with characterization of the eukaryotic and the archaeal DNA replicative proteins. This has led to a fairly good understanding of the bacterial replicative machinery, i.e. *E. coli*. Although, *E. coli* has been a good model for understanding the eukaryotic replicative machinery, there are still differences at molecular level that should be taken into account to better understanding the DNA replication process in eukaryotes.

Evolutionary studies and comparative genomics have suggested parallels between archaeal and eukaryotic replicative machinery [Edgell and Doolittle, 1997]. Yet, functional and structural studies in archaea suggest a striking homology between the archaeal and eukaryotic replicative machineries [Barry and Bell, 2006; Lindås and Bernander, 2013]. These observations suggest that archaea can be used as a model for gaining insight into elucidating the biochemical and structural proprieties of protein involved in DNA replication in eukaryotes. Taking into account the homology of the replisomes and principally the simplicity of the archaeal replicative machinery these organisms could provide a plethora of useful information about the replication of DNA in eukaryotic cells.

In eukaryotes, DNA replication begins with the recruitment of initiator proteins at the origin of replication, which leads to the formation of the pre-replicative complex (pre–RC) [Kornberg and Baker, 1992]. In particular, this process begins with

the recruitment, at the origin of replication, of ORC (Origin Recognition Complex), which in turn recruits Cdc6 (Cell division cycle 6) and Cdt1 (Chromatin licensing and DNA replication factor 1). These initiator factors are needed in order to recruit and load the heterotrimeric MCM2-7 complex (Minichromosome maintenance 2-7). MCM2-7 is an 3'→5' ATP–dependent helicase, which is needed for origin unwinding [Takeda and Dutta, 2005; Labib et al., 2001; Labib, 2000]. MCM2-7 also catalyses, along with Cdc45 (Cell division cycle protein 45) and the GINS complex (Go–ichi–ni–san, 5–1–2–3 in Japanese), the unwinding of DNA during the S phase of DNA replication. The Cdc45–MCM2-7–GINS complex is also known as CMG complex or '*unwindosome*' [Gambus et al., 2006; Moyer et al., 2006; Pacek et al., 2006]. The final step in replication initiation is the recruitment and loading of the replicative polymerases, which will carry out the synthesis of the new DNA strands [Takeda and Dutta, 2005]. Due to the antiparallel configuration of the DNA duplex, during chromosomal DNA replication the leading strand is replicated continuously while the lagging strand is extended discontinuously in the opposite direction, respect to the leading strand, by short RNA–DNA fragments, the Okazaki fragments [Kornberg and Baker, 1992]. Okazaki fragments are generated through strand displacement by the replicative polymerase. This displacement generates a specific DNA single strand called flap. In this process structure specific flap endonuclease 1 (Fen1) recognises the DNA flap and cleaves it leaving a nicked duplex DNA. Ligase I recognises this nick and seal it off [Kornberg and Baker, 1992]. Okazaki fragment maturation represents a required procedure to impart continuity to the lagging strand [Kornberg and Baker, 1992].

The first aim of my PhD is to gain insight into elucidating the biochemical properties of MCM protein complex from *Pyrococcus abyssi*. In particular, my aim is to investigate the mechanism by which the MCM is assembled at the origin of replication; the conformational change that the MCM helicase undergoes during the loading process at the origin of replication; the molecular mechanism by which the directionality of the unwinding is chosen and the mechanism by which MCM discriminate between ssDNA and dsDNA by using as model *P. abyssi*. *P. abyssi* was chosen as model since its single origin of replication is known [Matsunaga et al.,

2001, 2007; Myllykallio et al., 2000].

The second aim of my PhD is to investigate the architecture of the Okazaki fragment maturation machinery. It is known that the crenarchaeon *Sulfolobus solfataricus* possesses a simplified toolset for DNA replication compared to Eukaryotes [Dionne et al., 2003]. Interestingly, *S. solfataricus* has a subset of the eukaryotic Okazaki fragment maturation factors, among which there are a heterotrimeric DNA sliding clamp, the proliferating cell nuclear antigen (*Sso*PCNA), the DNA polymerase B1 (*Sso*PolB1), the flap endonuclease (*Sso*Fen1) and the ATP-dependent DNA ligase I (*Sso*LigI). *Sso*PCNA has been demonstrated to function as a scaffold with each subunit having a specific binding affinity for each of the factors involved in Okazaki fragment maturation [Dionne et al., 2003], and the most efficient coupling of activities occurs when a single *Sso*PCNA ring organises *Sso*PolB1, *Sso*Fen1 and *Sso* LigI into a complex Beattie and Bell [2012].

In my work, I will be using electron microscopy (EM) and single-particle analysis (SPA), which is a wonderful tool to perform structural studies on large macromolecular protein complexes.

# Chapter 3

# Materials and Methods

## 3.1 General microbiology

The bacterial strains used in this study are described in Table 3.2. *E. coli* XL-1 blue was used as host strain for plasmid propagation. Rosetta$^{TM}$(DE3) pLysS was used as host strain for proteins over-expression. Rosetta$^{TM}$(DE3) pLysS host strain are BL-21 derivatives designed to enhance the expression of proteins that contain codons rarely used in *E. coli*. These strains supply tRNAs for AGG, AGA, AUA, CUA, CCC, GGA codons on a compatible chloramphenicol-resistant plasmid.

Cells were grown and propagated in LB broth [Sambrook et al., 2001](Table 3.1). SOC medium was used to grow chemically competent cells after transformation by heat-shock methods [Sambrook et al., 2001]. Protein over-expression was carried out either in LB or TB. Antibiotics were added as selectable marker for either propagating, vectors and constructs, and during protein over-expression (Table 3.3).

## 3.2 Recombinant DNA

Plasmid vectors used in this study are described in Table 3.4. Plasmid DNA preparation, PCR products purification, gel extraction and mutagenesis were performed using commercially available kits (see Table 3.5) following the manufacturer's instructions. DNA was eluted in 70 $\mu$l of Milli–Q water and stored at -20°C.

**Table 3.1: List of media used in this study.**

| Buffer/Media | Ingredients |
|---|---|
| **Luria broth (LB)**[a] | 1 L |
| Bacto-tryptone | 10 gr |
| Bacto-yaest extract | 5 gr |
| NaCl | 10 gr |
| distilled $H_2O$ | to 1 L |
| pH was adjusted to 7.5 with NaOH | |
| **Luria agar (TA)**[a] | 1 L |
| Bacto-agar | 15 gr |
| Luria broth | to 1 L |
| **Terrific broth (TB)**[a] | 1 L |
| Bacto-tryptone | 12 gr |
| Bacto-yeast extract | 24 gr |
| Glycerol | 4 ml |
| 0.17 M $KH_2PO_4$ | 2.31 gr |
| 0.72 M $K_2HPO_4$ | 12.54 gr |
| distilled $H_2O$ | to 1 L |
| **Super optimal broth (SOB)**[a] | 1 L |
| Bacto-tryptone | 20 gr |
| Bacto-yeast extract | 5 gr |
| 10 mM NaCl | 0.58 gr |
| 2.5 mM KCl | 0.18 gr |
| 10 mM $MgCl_2$ | 0.95 gr |
| 20 mM glucose | 3.6 gr |
| distilled $H_2O$ | to 1 L |
| pH was adjusted to 7 with NaOH | |
| **Super optimal broth (SOC)**[a] **with Catabolite repression** | 1 L |
| 20 mM glucose | 3.6 gr |
| SOB | to 1 L |
| **Inoue trasformation buffer**[b] | 1 L |
| 55 mM $MnCl_2 \bullet 4H_2O$ | 10.88 gr |
| 15 mM $CaCl_2 \bullet 2H_2O$ | 2.2 gr |
| 250 mM KCl | 18.6 gr |
| 10 mM $MgCl_2$ | 0.95 gr |
| 10 mM PIPES pH 6.7 | 20 ml (0.5 M PIPES) |
| distilled $H_2O$ | to 1 L |

[a] Solutions were sterilised by autoclaving.

[b] Solution sterilised by filtration through a 0.22 $\mu$m Whatman® cellulose nitrate membrane filter.
**Note:** 2.31 g of $KH_2PO_4$ and 12.54 g of 0.72 M $K_2HPO_4$ were dissolved in 90 mL of $H_2O$.

**Table 3.2: Bacterial strains used in this study.**

| **Bacterial strain** | **Genotype** |
|---|---|
| XL1 Blue | endA1 gyrA96($nal^R$) thi − 1 recA1 relA1 lac glnV44 F′[:: Tn10proAB$^+$ $lacI^q$ Δ(lacZ)M15] hsdR17($r_K^- m_K^+$) |
| Rosetta$^{TM}$(DE3) pLysS | B F ompT hsdS($r_B m_B$) $dcm^+$ $Tet^R$gal γ(DE3)endA Hte [argU proL $Cam^R$][argU ileY leuW Strep/$Spec^R$] |
| BL21 CodonPlus$^{TM}$(DE3) RIPL | B F ompT hsdS($r_B m_B$) $dcm^+$ $Tet^R$gal γ(DE3)endA Hte [argU proL $Cam^R$][argU ileY leuW Strep/$Spec^R$] |

**Table 3.3: List of antibiotics used in this study.**

| **Antibiotic** | **Abbr.** | **μg/ml** | **Solvent** |
|---|---|---|---|
| Ampicillin | Amp | 100 | Water |
| Chloramphenicol | Chl | 34 | Ethanol |
| Kanamycin | Kan | 25 | Water |
| Tetracyclin | Tcn | 50 | Ethanol |

μg/ml, working concentrations.

**Table 3.4: List of plasmids used in this study.**

| Plasmid | Description | Tag | Selectable marker | Ref. |
|---|---|---|---|---|
| pTWO–E | Modified expression vector | 6–His[a] | Amp | none |
| pMAL$^{TM}$ | Expression vector | MBP[b] | Amp | [Miller et al., 2001] |
| pET3a–Tr | Modified polycistronic expression vector | none | Amp | [Tan, 2001] |

[a] MBP, maltose binding protein tag.

[b] 6–His, six histidine tag.

**Table 3.5: List of kits used in this study.**

| Kit | Comapny |
|---|---|
| QIAprep® spin miniprep kit | QUIAGEN |
| Qiaquick® PCR purification kit | QUIAGEN |
| Qiaquick® gel extraction kit | QUIAGEN |
| Quick Ligation$^{TM}$ | New Englad Biolab |
| QuickChange® | Agilent |

**Table 3.6: Typical cycle programe.**

| | | |
|---|---|---|
| Initial template denaturation: | 95°C | 5 min |
| | 30–40x cycles | |
| Template denaturation: | 95°C | 30 sec |
| Primers annealing: | 45–65°C | 30 sec |
| Extension: | 72°C | 30 sec–1min |
| Final extension: | 72°C | 10 min |
| Storage: | 10°C | $\infty$ |

**Table 3.7: List of oligos used in this study.**

| Use | Primer | Sequence 5'-3' | RS | Length | Tm° |
|---|---|---|---|---|---|
| PCR and sequencing | T7 | TAATACGACTCACTATAGGG | ndt | 20 | 50.9° |
| | STO720 | TGTGAAATTGTTATCCGCT | ndt | 19 | 60.2° |
| | F_PAB2373–1 | TATATACATATGGATAGAGAGGAGATCATCGAGAGATTCCTG | NdeI | 42 | 73° |
| | R_PAB2373 | GCGCGGGGTACCTCAGACGGTTCTGTAATAACC | KpnI | 33 | 79° |
| | F_PAB2373_in | P–ACGGCCGCGGTGG | ndt | 13 | 43° |
| | R_PAB2373_in | P–TTTCGCGACTCCCGG | ndt | 15 | 43° |
| | J–NTERM | GATCCGGGAGTCGCGAAAAGCCAACTTCTCAGATAC | ndt | 36 | 81° |
| | J–CTERM | CTAACCACCGCGGCCGTGAGCCCAGCGGCAGAACTC | ndt | 36 | 90° |
| | F_PAB0956 | TATATACATATGGATAGAGAGGAGAT CATCGAGAGATTCCTG | | 42 | 73° |
| | R_PAB0956 | GCGCGGGGTACCTCAGACGGTTCTG TAATAACC | | 33 | 79° |
| | F_PAB1566 | TATATACATATGGATAGAGAGGAGAT CATCGAGAGATTCCTG | | 42 | 73° |
| | R_PAB1566 | GCGCGGGGTACCTCAGACGGTTCTG TAATAACC | | 33 | 79° |
| DNA bidning assay | oligo3 | CAAATTCCCATATGTGGCGCGCCGTGTATACGACTCCCTGCAG GAAACCATGGCATCCG | ndt | 59 | 72.5° |
| | oligo4 | CGGATGCCATGGTTTCCTGCAGGGAGTCGTATACACGGCGCGC CACATATGGGAATTTG | ndt | 59 | 72.5° |
| Okazakisome reconstruction | Template | TTAAAGTTAGGTGGGGGACTCTGCCTCAAGACGGTAGTCAACG TGACCGCAGCAAACCTG | ndt | 60 | 92° |
| | Upstream | CAGGTGGGCTGCGGTCACGTTGACTAGGGTCG | ndt | 32 | 86.3° |
| | Downstream | CAAGCAGTCCTAACTTTGAGGCAGAGTCCCCCACCTAACTTTAA | ndt | 44 | 80.6° |

**Note:** RS, Restriction site

### 3.2.1 Agarose gel electrophoresis of DNA

DNA fragments from PCR products, plasmid DNA and linearised plasmid were separated on either 1% or 1.5% w/v agarose gel, whereas purification of DNA fragments was performed by running samples on 0.8% w/v agarose gel (see Table 3.15). The appropriate amount of agarose was dissolved in 1x TAE and boiled at 100°C til the agarose was completely dissolved. The solution was allowed to cool to 55°C. Ethidium bromide (final concentration of 0.5 $\mu$g/ml) was added to visualise DNA under UV light using a Gel–Doc$^{TM}$ system (BioRad). Gels were run at 50V in 1x TAE buffer up to. The size of fragments was checked by comparison with a DNA ladder (Table 3.15). DNA was quantified using a Nanodrop (ND-100 v3.5).

### 3.2.2 Polymerase chain reaction (PCR)

Pwo DNA polymerase is a highly processive 5'–3' DNA polymerase with proofreading activity. Because of these qualities, it was chosen as polymerase for all the PCR reactions when the product was required for cloning. When PCR reactions were carried out for routine checking of fragment sizes, for example colony PCR, Promega GoTaq® DNA polymerase was used instead. Master mixes for PCR amplification were prepared according to the manufacturer's instruction. Primer annealing temperatures were dependent on primer sequence and were altered accordingly (Table 3.7). The extension time was determined by the polymerase of choice and the length of the template to be amplified. A typical cycle program is shown in Table 3.6. Reactions were carried out using the Eppendorf Mastercycler Gradient PCR machine.

**Colony PCR** Colony PCR screening is a very quick PCR–based method commonly employed for screening recombinant colonies after transformation. Colony PCR can be effectively used to identify recombinant clones as well as identify insertions and deletions; to determine the orientation of a DNA insertion and to directly amplify a desired DNA fragment [Woodman, 2008]. When colony PCR was performed, small amounts of individual bacterial colonies were removed from the plate and resuspended into 10 $\mu$l of Milli–Q water. 2 $\mu$l of this suspension was directly

placed into PCR tubes with 18 $\mu$l of master mix.

**Cross–over PCR** Crossover PCR (Figure 3.1) was used in order to join two separate fragments of DNA without the need for restriction and ligation. For this experiment purpose, four primers were required, two for amplifying the first DNA fragment and two for amplifying the second DNA fragment. Internal primers, the reverse primer for the first DNA fragment and the forward primer for the second DNA fragment, were designed to have 20–25 base pairs of homology to each other. Initially, the two DNA fragments were amplified separately, creating two products that at one extremity contained 20–25 homologous base pairs. Finally, to join the two DNA products, a PCR was set up where the two products from the initial PCR reactions were used as template. During melting and cooling of the templates, the region of homology would bring the two DNA fragments together to create a single new fragment for amplification using the external primers, the forward primer for the first DNA fragment and the reverse primer for the second DNA fragment. The resulting PCR product would be a fusion of the two DNA fragments of interest. This method was successfully used to reconstruct *Pab*MCM full-length.

**PCR–based site–directed mutagenesis** In vitro site–directed mutagenesis is a technique employed for carrying out modification at level of nucleotide sequence. For this purpose a commercially available kit (Table 3.5) was used for engineering point mutations in order to modify the nucleotide sequence of the predicted secondary RBS found in the *Pab*MCM ORF. Manufacturer's instructions were followed.

### 3.2.3 Restriction digestion and ligation of purified DNA fragments

Restriction enzymes and buffers were obtained from New England Biolabs (NEB). 1–2 $\mu$g of either PCR purified DNA or plasmid DNA were digested following manufacturer's instructions. Samples were incubated at the optimum digestion temperature (37°C) for 2-4 hours. Digested plasmid DNA was purified onto 0.8% agarose gel. Ligation reactions were performed following manufacturer's instructions.

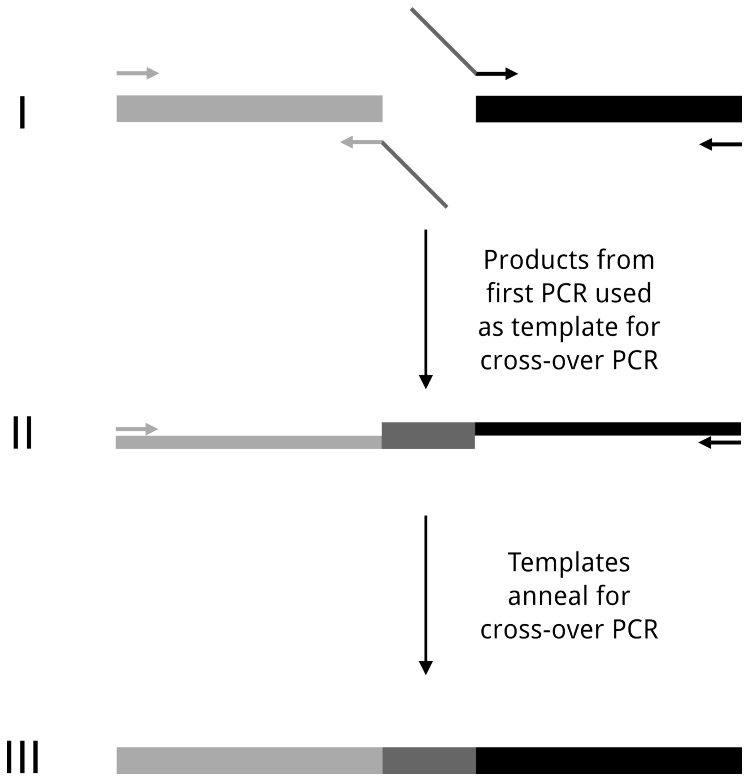



**Figure 3.1: Crossover PCR.** (I) Initially, the two DNA fragments (shown in light grey and black) are amplified separately using primers. Tailed–primers are designed to have a 20-25 bp region of homology to each other. (II) In the crossover PCR, the products from the first PCR reaction are used as template. When these melt and reanneal, the region of homology between them will bring them together. (III) This will create a new, single, template for amplification using primers.

### 3.2.4 Preparation of chemically competent *E. coli* cells

Chemically competent *E. coli* cells were prepared as in Sambrook and Russell [2006]. This protocol generates "ultra–competent" cells that yield 1 x $10^8$ to 3 x $10^8$ transformed colonies per $\mu$g of plasmid DNA. Briefly, an overnight culture of the strain to be transformed was grown in SOB medium. This culture was used as starter culture to inoculate three 1–liter flasks, each containing 250 ml of SOB. The first flask received 10 ml of starter culture, the second received 4 ml, and the third received 2 ml. All three flasks were incubated overnight at 18–22°C with moderate shaking. The $OD_{600}$ was monitored each 45 min. When the $OD_{600}$ of one of the cultures reached 0.55, the culture vessel was transferred to an ice–water bath for 10 minutes. The two other cultures were discarded. Cells were harvested by centrifugation at 2500g in a Sorvall GSA for 10 minutes at 4°C. Supernatant was discarded and cells were gently resuspended in 80 ml of ice–cold Inoue transformation buffer. Once again, cells were harvested by centrifugation at 2500g in a Sorvall GSA for 10 minutes at 4°C. Supernatant was discarded and cells were gently resuspended in 20 ml of ice–cold Inoue transformation buffer. 1.5 ml of DMSO was added and the bacterial suspension was mixed by swirling. Finally, the bacterial suspension was dispensed in 50 $\mu$l aliquots in sterile microcentrifuge tubes and immediately snap–frozen by immersing the tube is a bath of liquid nitrogen. Tubes were stored at –80°C until needed.

### 3.2.5 Transformation of chemically competent *E. coli* cells by heat–shock

Transformation was carried by heat–shock. Briefly, the transforming DNA (20–100 ng) was mixed with 50 $\mu$l of thawed competent cells and incubated on ice for 30 minutes. Cells were heat–shocked in a water bath at 42°C. The duration of the heat-shock step varied depending upon the bacterial strain (45" for XL1 blue whereas 30" for both BL21 CodonPlus$^{TM}$(DE3) RIPL and Rosetta$^{TM}$(DE3) pLysS). After heat–shock, cells were placed on ice for 15 minutes. 950 $\mu$l of SOB were added and cells were incubated at 37°C with vigorous shacking for $\sim$1 hour in order to recover.

Finally, 100 $\mu$l of bacterial transformation was plated on LB agar plates with the appropriate antibiotic. For each transformation, a negative control devoided of DNA was prepared in parallel.

### 3.2.6 Gene sequencing

Sequencing reactions were carried out at the GenePool, University of Edinburgh (http://genepool.bio.ed.ac.uk/). The sequencing reactions were prepared by mixing 5$\mu$l of of DNA (∼500ng) with 1$\mu$l of T7 promoter primer for forward sequencing reaction as STO720 was used for reverse sequencing reactions. Both primers were at final concentration of 3.2 pmole. Reads were manually analysed by checking the chromatogram using FinchTV software (http://www.geospiza.com/Products/finchtv.shtml).

### 3.2.7 Bioinformatics

Sequences analysis and manipulation was carried out with Expasy (SIB Bioinformatics Resource Portal) at http://www.expasy.org/. DNA sequences were fetched from National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/), whereas protein sequences from Uniprot databank (http://www.uniprot.org/). Protein data bank (PDB) (http://www.rcsb.org/pdb/home/home.do) and Electron microscopy data bank (EMDB) (http://www.ebi.ac.uk/pdbe/emdb/) were used for retrieving structural information.

## 3.3 Biochemical protein characterization

### 3.3.1 SDS–PAGE

Denaturing protein gel electrophoresis was carried out using the Tris–glycine buffer system described previously [Laemmli et al., 1970]. A list of buffers for SDS–PAGE is provided in Table 3.8.

The gel matrix was produced by co–polymerizing acrylamide and bis-acrylamide using TEMED as a crosslinker and APS as starter catalyst. The gel consisted of

Table 3.8: List of buffers for SDS–PAGE.

| Buffer/Media | Ingredients |
|---|---|
| **10% Ammonium persulfate (APS)** | 10 ml |
| 10% Ammonium persulfate(w/v) | 1 gr |
| distilled $H_2O$ | to 10 ml |
| **6x Loading sample buffer** | 10 ml |
| 0.37 M Tris | 0.59 gr |
| 48% glycerol (w/v) | 4.8 ml (100% glycerol) |
| 6% SDS (pellet) (w/v) | 0.6 gr |
| 9% $\beta$-mercaptoethanol | 0.9 ml (14.7 M $\beta$-mercaptoethanol) |
| 0.03% Bromophenol blue | 3 mg |
| distilled $H_2O$ | to 10 ml |
| pH was adjusted to 6.8 with HCl | |
| **5% Stacking gel** | 10 ml |
| 5% Acrylamide (v/v) | 1.7 ml 30% Acrylamide |
| 125 mM Tris pH 6.8 | 1.25 ml (Tris-HCl pH 6.8) |
| 1% SDS (v/v) | 0.1 ml (20% SDS) |
| 1% APS (v/v) | 0.1 ml (10 APS) |
| 0.1% TEMED (v/v) | 0.01 ml TEMED |
| distilled $H_2O$ | to 10 ml |
| **12% Stacking gel** | 50 ml |
| 12% Acrylamide (v/v) | 20 ml 30% Acrylamide |
| 250 mM Tris pH 8.8 | 12.5 ml (Tris-HCl pH 8.8) |
| 1% SDS (v/v) | 0.5 ml (20% SDS) |
| 1% APS (v/v) | 0.5 ml (10 APS) |
| 0.1% TEMED (v/v) | 0.05 ml TEMED |
| distilled $H_2O$ | to 50 ml |
| **15% Stacking gel** | 50 ml |
| 15% Acrylamide (v/v) | 25 ml 30% Acrylamide |
| 250 mM Tris pH 8.8 | 12.5 ml (Tris-HCl pH 8.8) |
| 1% SDS (v/v) | 0.5 ml (20% SDS) |
| 1% APS (v/v) | 0.5 ml (10 APS) |
| 0.1% TEMED (v/v) | 0.05 ml TEMED |
| distilled $H_2O$ | to 50 ml |

two parts; a wide meshed stacking gel with a pH of 6.8 and an higher concentration resolving gel at pH 8.8. This pH–shift leads to an acceleration of the glycine molecules (due to a change from its almost unloaded to a strongly ionic form) that now overtake the proteins causing them to run in a discrete band. Once the casting chamber was assembled and the resolving gel cast, 2–propanol was overlaid directly after pouring the resolving gel in order to avoid an oxidation and dehydration of the surface and remove eventual bubbles formed after pouring the resolving gel. After the gel had polymerized, the 2–propanol was thoroughly washed off and the stacking gel was applied on the top of the resolving gel. A comb was added to the stacking gel in order to form the wells for loading samples. After polymerisation was complete, the gels were run vertically (PerfectBlue Dual Gel System, PEQLAB or Mini–Protean II®, BIORAD) in the conventional discontinuous Running Buffer system at constant current of 30 mA.

Cell were harvested and resuspended in a final volume of 6x sample buffer according to the following formula:

$$V_f = \frac{OD_{600}100}{1.5}$$

Where

$V_f$: final volume of sample buffer.

$OD_{600}$: Cell $OD_{600}$ measured.

Samples were boiled at 95°C for 2 minutes prior to loading. Molecular Mass Standards (pre–stained, INVITROGEN) served to determine the apparent molecular mass in SDS–PAGE of the separated proteins.

### Comassie blue staining

Commassie blue staining solution was prepared by dissolving first the Comassie brilliant blue in methanol and then adding glacial acetic acid and water. Quantity and volumes accordingly to Table 3.9. For the protein gel staining, gel was placed in a plastic box and a volume, so to cover the gel, was added to it. Gels were stained for an hour and destianed for overnight in destaining solution (Table 3.9).

**Table 3.9: List of buffers for protein gel staining.**

| Stain | Ingredients |
|---|---|
| **Comassie staining** | |
| **Comassie R–250 staining solution** | 1L |
| Comassie brilliant blue R–250 | 1 gr |
| 50% Methanol (v/v) | 500 ml |
| 10% Glacial acetic acid (v/v) | 100 ml |
| distilled $H_2O$ | to 1 L |
| **Comassie destaining solution** | 1 L |
| 30% Methanol (v/v) | 300 ml |
| 10% Glacial acetic acid (v/v) | 100 ml |
| distilled $H_2O$ | to 1 L |
| **Silver stain** | |
| **Fixing solution I** | 50 ml |
| 50% Methanol (v/v) | 25 ml |
| 10% Glacial acetic acid (v/v) | 5 ml |
| Ultra–pure $H_2O$ | to 50 ml |
| **Fixing solution II** | 50 ml |
| 20% Ethanol (v/v) | 10 ml |
| Ultra–pure $H_2O$ | to 50 ml |
| **Sensitising solution** | 50 ml |
| 0.02% Sodium thiosulpahte (v/v) | 0.1 ml 10%(w/v)[ss] |
| Ultra–pure $H_2O$ | to 50 ml |
| **Silver nitrate solution** | 50 ml |
| 0.2% Silver nitrate (v/v) | 0.1 ml 10%(w/v)[ss] |
| Ultra–pure $H_2O$ | to 50 ml |
| **Developing solution** | 50 ml |
| Formaldehyde | 0.0125 ml |
| Potassium carbonate | 3 ml |
| Sodium thiosulpahte | 0.006 ml |
| Ultra–pure $H_2O$ | to 50 ml |
| **Stop solution** | 50 ml |
| 4% Tris–base (w/v) | 2 gr |
| 2% Acetic acid (v/v) | 1 ml |
| Ultra–pure $H_2O$ | to 50 ml |

[ss] Stock solution as in Table 3.14

**Silver staining**

The silver staining protocol used in this study was an optimised, by the author of this study, shoerter version of the A protocol in [Chevallet et al., 2006]. Protcol's steps as following:

- Incubate gel in fixing solution I for 15 minutes on shaking.
- Incubate gel in fixing solution II for 10 minutes on shaking.
- Incubate gel in ultra-pure water for 10 minutes on shaking.
- Incubate gel in sensitising solution for 2 minutes on shaking.
- Wash gel in ultra-pure water for 1 minute.
- Incubate gel in silver nitrate solution for 10 minutes.
- Wash in ultra-pure water for 1 minute.
- Develop staining in developing solution until a good homogeneus stain is visualised.
- Stop staining developing by discarding it and adding stop solution.
- Store gel in stop solution for no more than a week.

After each step mentioned above, solution were discarded. Silver staining was carried out in plastic boxes carefully washed with acetone before use. Gels were manipulated with acetone–cleaned spatula since touching gels with gloves leaves a print on gels.

### 3.3.2 Western blotting analysis

Antibodies against *Sso*PCNA123, *Sso*PolB1, *Sso*LigI and *Sso*Fen1 were provided by Prof. Stephen D Bell; whereas antibodies against 6–Histidine tag were purchased from Sigma (see Table 3.15). Samples were prepared and ran onto SDS–PAGE gels as described before (see SDS–PAGE). The SDS–PAGE gel was transferred to a nitrocellulose membrane accordingly to the manufacturer's instruction. After transfer,

**Table 3.10: List of buffer used for western blotting analysis.**

| **Buffer** | **Ingredients** |
|---|---|
| **10X Tris buffer Saline (TBS)** | 1L |
| 0.5 M Tris | 60.5 gr |
| 1.5 M NaCl | 87.6 gr |
| distilled $H_2O$ | to 1 L |
| **10X Tris buffer Saline tween20 (TBST)** | 1L |
| 0.5 M Tris | 242 gr |
| 1.5 M NaCl | 57.1 ml |
| 10% Tween20 | 1 ml |
| distilled $H_2O$ | to 1 L |
| pH was adjusted to 7.5 with HCl | |
| **10x Transfer buffer** | 1L |
| 0.25 M Tris | 30 gr |
| 1.9 M glycine | 144 gr |
| distilled $H_2O$ | to 1 L |
| **Blocking solution** | 200 ml |
| 5% Skim milk (w/v) | 10 gr |
| 1X TBST | to 200 ml |

the membrane was rinsed twice in TBST and incubated in blocking solution for 1 hour. Primary antibodies, for the protein being immunodetected, were incubated at different dilutions ($\alpha$–*Sso*PCNA123, 1:$10^4$; $\alpha$–*Sso*PolB1, 1:$10^3$; $\alpha$–*Sso*LigI, 1:$10^3$; $\alpha$–*Sso*Fen1, 1:$10^3$ and $\alpha$–6His–tag, 1: $10^4$) in blocking solution for 1 hour. Membranes were rinsed twice in TBST and incubated in the presence of HRP–conjugated secondary antibodies (see Table 3.15) in 5% milk in TBST for 1 hour. Two washes in TBST, followed by two more washes in TBS were performed. Chemiluminescent detection was carried out with SuperSignal Western Blotting Kits whereas colorimetric one was carried out with SigmaFast$^{TM}$BCIP®/NBT (see Table 3.15) accordingly to manufacturer's instructions. A list of buffers used for immunodetection is provided in Table 3.10.

## 3.4 Protein purification

Protein purification was carried out using a combination of heat denaturation and chromatographic techniques including affinity–tag, strong ion exchange and size exclusion chromatography. In Table 3.11 are summed up the steps carried out during purification of *Pab*MCM, *Sso*PCNA123, *Sso*PolB1, *Sso*LigI and *Sso*Fen1.

### 3.4.1 Cell lysis

Recombinant proteins overexpressed in Rosetta$^{TM}$(DE3) pLysS were harvested by centrifugation at 3000 x g. Cell disruption was achieved by either sonication or cells grounding in liquid nitrogen with the aid of a mortar. Cell disruption by sonication was performed at 50% amplitude. To avoid degradation from bacterial proteases, protease inhibitors were added prior to cell disruption. Separation of soluble proteins from the insoluble fraction (cell debris) was achieved by centrifugation (details of cell lysis in Table 3.11).

**Table 3.11: Protein purification protocols.**

| | Recombinat proteins | | | | |
|---|---|---|---|---|---|
| | **PabMCM** | **SsoPCNA123** | **SsoPolB1** | **SsoLigI** | **SsoFen1** |
| Lysis buffer | 20 mM HEPES pH 7.4, 500 mM NaCl + PIC | 20 mM HEPES pH 8.0, 300 mM NaCl + PIC | 10 mM HEPES pH 7.5, 100 mM NaCl, 1mM DTT + PIC | 10 mM HEPES pH 8.0, 300 mM NaCl + PIC | 20 mM MES pH 6.0, 300 mM NaCl + PIC |
| Cell distruption | Sonication | Sonication | Sonication | Grounding in liquid $N_2$ | Sonication |
| Heat denaturation | 20 minutes at 70°C | 25 minutes at 75°C | 20 minutes at 65°C | 25 minutes at 75°C | 15 minutes at 75°C |
| Clarification | 22000 x g, 20 minutes | 22000 x g, 20 minutes | 22000 x g, 20 minutes | 22000 x g, 20 minutes | 22000 x g, 20 minutes |
| $1^{st}$ Chromatographic step | 1 ml Ni–NTA HiTrap | 1 ml Ni–NTA HiTrap | 1 ml HiTrap Heparin | 1 ml Ni–NTA HiTrap | 1 ml HiTrap Heparin |
| Sample dilution | | | | | 3–fold |
| Column equilibration | 20 mM HEPES pH 8.0, 500 mM NaCl | 20 mM HEPES pH 8.0, 300 mM NaCl | 10 mM HEPES pH 7.5, 100 mM NaCl, 1 mM DTT | | 20 mM MES pH 6.0, 30 mM NaCl, 1 mM EDTA, 0.5 DTT |
| Elution buffer (B%) | 20 mM HEPES pH 8.0, 500 mM NaCl, 500 mM imidazole | 20 mM HEPES pH 8.0, 300 mM NaCl, 500 mM imidazole | 10 mM HEPES pH 7.5, 1000 mM NaCl, 1 mM DTT | | 20 mM MES pH 6.0, 1000 mM NaCl, 1 mM EDTA, 0.5 DTT |
| Elution gradient | 15 CV | 15 CV | 75 CV | | 75 CV |
| $2^{nd}$ Chromatographic step | SEC in 20 mM HEPES pH 7.4, 150 mM NaCl | SEC in 20 mM HEPES pH 8.0, 300 mM NaCl | SEC in 20 mM HEPES pH 8.0, 500 mM NaCl | SEC in 20 mM HEPES pH 8.0, 150 mM NaCl, 14 mM $\beta$–mercaptoethanol | 20 mM MES pH 6.0, 150 mM NaCl, 1 mM EDTA, 0.5 DTT |
| $3^{rd}$ Chromatographic step | | | 1 ml HiTrap SP | 1 ml HiTrap Q | |
| Sample dilution | | | 5–fold | 2–fold | |
| Column equilibrated | | | 20 mM HEPES pH 7.5, 100 mM NaCl, 1 mM DTT | 20 mM HEPES pH 8.0, 75 mM NaCl, 14 mM $\beta$–mercaptoethanol | |
| Elution buffer (B%) | | | 20 mM HEPES pH 7.5, 1000 mM NaCl, 1 mM DTT | 20 mM HEPES pH 8.0, 1000 mM NaCl, 14 mM $\beta$–mercaptoethanol | |
| Elution gradient | | | 15 CV | 15 CV | |
| Storage temperature | –80°C | –80°C | –80°C | –80°C | –80°C |

### 3.4.2 Heat denaturation step

The thermostable nature of hyperthermophilic archaea allows the inclusion of a heat denaturation step in the protein purification protocol. During this step, a good amount of *E. coli* proteins were denatured, resulting in a reduced amount of contaminants in the cell lysate (details of heat denaturation in Table 3.11).

### 3.4.3 Chromatographic techniques

Proteins were purified using chromatography techniques that separated them according to differences in their specific properties. Chromatographic techniques mainly used in this study are discussed.

#### Affinity chromatography

Affinity chromatographic is a special chromatographic technique, which is based on the biorecognition between two biomolecules, such as interactions between enzymes and substrates, receptors and ligands, or antibodies and antigens. This specific interaction, typically reversible, enable recombinant proteins to be purified quite straightforwardly from a complex mixture, which in this specific case is the pool of the host proteins. The affinity ligand is covalently bound onto a solid matrix to create a stationary phase whereas the target molecule (tag) is in the mobile phase [Urh et al., 2009]. In this study, immobilized-metal affinity chromatography (IMAC) was used. IMAC is a separation technique that uses covalently bound chelating compounds on a solid matrix to trap metal ions (affinity ligand). IMAC widely used especially when rapid purification and substantial purity of the product are necessary. However compared to other affinity separation technologies it cannot be classified as highly specific, but only moderately so. On the other hand, IMAC offers a number of advantages over specific affinity chromatographic techniques. The benefits of IMAC are: ligand stability; high protein loading; mild elution conditions and simple regeneration and low cost [Gaberc-Porekar and Menart, 2001].

In this study, HiTrap IMAC (GE) columns, nickel charged, were used in order to perform IMAC. Elution was performed with 500 mM imidazole (details of IMAC

purification in Table 3.11).

### Ion exchange chromatography (IEX)

Proteins are charged biomolecules. Their net charge depends on particular amino acids with ionizable groups. Since all molecules with ionizable groups can be titrated, their net charge is highly dependent on the pH of the buffer in which they are solubilised. IEX separates biomolecules on the basis of differences in their net charge surface. More specifically, IEX takes advantage of the unique relationship, which is specific for each protein, between net surface charge of and pH. In IEX, charged molecules bind onto an oppositely charged matrix. Consequently, a protein, which has no net charge at a pH equivalent to its isoelectric point (pI), will not bind to a charged matrix. At pH above its pI, the net charge is negative and thus the protein will bind to a positively charged matrix in which case it is known as cation exchange chromatography. At pH below its pI, the net charge is positive and thus the protein will bind to negatively charged, in which case it is known as anion exchange chromatography. Elution of bound proteins is carried out either increasing the ionic strength or changing the pH of buffer [Alois and Rainer, 2009]. In this study, 1 mL FastFlow Q columns (GE), for cation exchange chromatography and 1 ml FastFlow S, for anion exchange, were used accordingly to the manufacturer's instruction. Columns were equilibrated with buffer the appropriate buffer. The salt concentration was kept at low levels to promote binding to the charged matrix. A washing step of 20–30 CV was used to remove non–specific binding proteins. Samples were eluted by a linear gradient (details of IEX purification in Table 3.11).

### Size exclusion chromatography (SEC)

Size exclusion chromatography, better know as gel filtration, is chromatographic technique which separates biomolecules according to the difference in molecular size. Gel filtration is performed using columns in which porous beads, the chromatographic support, are packed into it. Unlike to other chromatographic techniques, biomolecules do not bind the chromatographic support (stationary phase) but they rather diffuse into it. Diffusion is inversely proportional to the MW, hence bigger

biomolecules diffuse slower into the porous matrix, resulting in a early elution from the column. In contrast, smaller biomolecules diffuse faster into the porous matrix, resulting in a late elution respect to the bigger ones. This differences in the capacity of diffuse into the pores of the matrix enable sepration of molecules based onto their MW [Stellwagen, 2009]. In this study, SEC was performed using either Superose$^{TM}$ 6 10/300 GL and Superdex$^{TM}$ 200 10/300 GL columns (GE) (details of SEC purification in Table 3.11).

## 3.5 GraFix method

GraFix (Gradient Fixation) is a method for sample preparation for single particle electron cryo–microscopy, in which a glycerol gradient centrifugation step is coupled with a gradient of glutaraldehyde. This method has been successfully used to either reduce heterogeneity of macromolecules or stabilise protein complexes, which present dissociation upon dilution [Kastner et al., 2007]. In this study, GraFix was used to stabilise the highly dynamic Okazaki fragment maturation protein complex.

The Grafix gradient was prepared manually with the aid of a gradient mixer connected to a peristaltic pump. The gradient mixer was a gift from Dr. Jim Allan (University of Edinburgh).

Solutions were prepared as shown in Table 3.12. 2 ml of solution H were poured into the cylinder directly connected to the peristaltic pump, whereas 2 ml of solution L were poured into the other cylinder. Cylinders were connected to each other by a valve. The gradient was formed into a 4 ml polyallomer tube (Beckman Cat. no. 328874) by starting the peristaltic pump while opening the valve that connected the two cylinders with H and L solution. Once the gradient was formed, 200 $\mu$l were removed from the top of the gradient. After then, 100 $\mu$l of 5% cushion were added on the top of the gradient (5% cushion was obtained by dilution of L solution). 200 $\mu$l of the Okazakisome reconstructed complex were added onto the 5% glycerol cushion. Tube were checked for balance and then loaded into the centrifuge bucket. The GraFix gradient was spun for 18 hours at 120.000 x g (34.000 rpm) at 4° using a Sorvall® THT60.4 swing out rotor placed into a Sorvall® ultracentrifuge.

**Table 3.12: List of solution used for GraFix method.**

| Buffer | Ingredients |
|---|---|
| **10X Buffer** | 10 ml |
| 100 mM HEPS pH 8.0 | 1 ml 1 M[SS] |
| 1.5 M NaCl | 3 ml[SS] |
| 50 mM $MgCl_2$ | 0.5 ml[SS] |
| Nuclease–free $H_2O$ | to 1 L |
| **Light Solution (L)** | 10 ml |
| 10% Glycerol (v/v) | 1.6 gr |
| 1X buffer | 1 ml 10X buffer |
| Nuclease–free $H_2O$ | to 10 ml |
| **Heavy Solution (H)** | 10 ml |
| 30% Glycerol (v/v) | 3.78 gr |
| 0.15% Glutaraldehyde | 0.06 ml of 25% stock |
| 1X buffer | 1 ml 10X buffer |
| Nuclease–free $H_2O$ | to 10 ml |

[SS] Stock Solution as in Table 3.14.

Fractionation was achieved manually by collecting 22 fractions of ∼180 $\mu$l each from the top of the gradient with an Hamilton syringe. Fractions where analysed by SDS–PAGE and western blotting.

## 3.6 DNA binding assay

DNA binding assay was carried out as in [Brewster et al., 2008]. Oligo3 and 4 were annealed in 1X annealing buffer to a final concentration of 10 $\mu$M. Annealing was achieved by incubating the oligos mixture in boiling water and letting it cool overnight.

Different concentrations of *Pab*MCM and DNA were tested in 20 mM HEPES, 150 mM NaCl. Reactions were carried out at 50°C for 30 minutes. Annealing was checked by running the annealed oligos onto 1.5 % agarose gel. Gels were run without any particular modification to the method described in section 3.2.1.

## 3.7 Electron microscopy

### 3.7.1 Grid preparation

#### Preparation of carbon coated grids

Microscopy of negatively stained samples was performed using home–made carbon coated grids. Copper grids (400 mesh, 3.05mm diameter) were purchased from Agar Scientific whereas ultra–thin carbon layer was prepared in house.

**Producing thin continuous carbon film** The carbon layer was produced with the aid of the Edwards coating system unit (model E3O6A). To evaporate carbon, two carbon roads were drilled as shown in Figure 3.2F and tips were polished by rubbing them on filter paper. The two carbon rods were fitted into the fixed carbon holder. The double–neck carbon rod was fitted into the movable carbon holder aligned and with the fixed, blunt end carbon rod. The carbon source holder was fixed into the power circuit of the machine. Mica sheets were placed on the bottom of the coater with the freshly split side up. The distance between the carbon source and

mica sheets was approximately 10–13 cm. The evaporation chamber was evacuated with high vacuum (1x$10^{-5}$ to 1x$10^{-6}$ Torr). When the vacuum was good enough, the carbon was evaporated with a current (low tension) and spread over the mica sheets. Covered in a Petri dish the carbon coated mica was rested for at least 24 h.

**Covering grids with thin carbon layer** Copper grids were covered with the carbon film as follows:

- Grids were placed (shiny side up) on a metal carrier net that was fixed in a water container so that it formed a plain parallel to and about 1 cm below the water surface.
- The carbon coated mica sheet (carbon upside) was inserted into the water in a 45°angle at a constant speed letting the carbon layer float off the mica.
- The water level was lowered and the carbon film was directed onto the grids with tweezers.
- Grids were ready for use after drying over night.

**Assessing carbon layer thickness** Carbon thickness was assessed by OD with a densitometer (Helland electronic Wetzlar) as shown in Agar [1957]. The thickness in nanometres is roughly equal to the OD measured times 100 (e.g. an OD of 0.05 is ∼5nm). Carbon film was normally evaporated to a thickness ranging from ∼5nm to ∼8nm.

### 3.7.2 Specimen preparation: negative stain with uranyl acetate

Just before applying the specimen, grids were air–glow discharged for 15 second in an EMtech K100x glow–discharge chamber unit, running at 3x10 mbar and 25 mAmp to render the grid surface hydrophilic, thereby enabling improved specimen adhesion. 4 $\mu$l of freshly purified protein sample (10–20 $\mu$g/ml) were applied onto a freshly glow–discharged grid and incubated for 2 minutes. The excess solution was

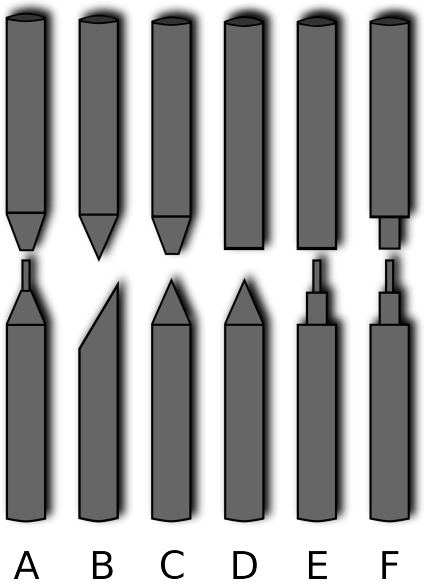



**Figure 3.2: Possible carbon pointed rod shapes.** Carbon rods can be either drilled or sharpened in different shapes. (A) Carbon rod with conical blunt sharpened end whereas its complementary carbon rod has a conical blunt sharpened end with a thinner tip point. (B) Carbon rods set up with one conical pointed tip whereas the complementary rod has an oblique blunt end. (C) Carbon rods set up with one conical blunt end tip and its complementary conical blunt end tip carbon rod. (D) Carbon rods set up with one blunt end carbon rod whereas the complementary carbon rod is blunt end. (E) Carbon rods set up with one blunt end carbon rod and its complementary carbon rods with a double–neck drilled end with a sharpened thinner tip point. (F) One carbon rod had a thinner cylindrical blunt drilled end whereas the complementary rod had a double–neck drilled end with the sharpened thinner tip point.

then removed by blotting the edge of the grid onto a filter paper (Whatman$^{TM}$No 1). The grid was then washed and stained by catching and blotting three drops ($\sim$20$\mu$l) of ultra–pure water and four drops ($\sim$10$\mu$l) of 2% (w/v) uranyl acetate. The last droplet was allowed to remain for five minutes on the grid before blotting. During this process the grid was mounted in a pair of forceps. Finally the grid was air dried before being stored at room temperature in a grid storage box.

## 3.8 Transmission electron microscopy (TEM)

Negatively stained grids were first viewed at the Philips CM120 electron microscope, operating at 80 kV and equipped with an Orius$^{TM}$ CCD camera (Gatan), for quality assessment of the protein samples. Because of the short heating time of the tungsten filament, the easy handling and the availability of a CCD camera with continuous readout mode, this method allowed analyses of a high number of grids in a short time. Samples were analyzed regarding their concentration, staining and homogeneity.

Data acquisition for final 3D reconstruction, was carried out using a FEI Tecnai$^{TM}$ F20 field emission gun (FEG) electron microscope equipped with TVIPS$^{TM}$ TemCam–F816 CMOS camera, which has 128 x 128 mm$^2$ sensor area with 64 megapixels resolution.

Before each data acquisition, electron bream, apertures and microscope lenses were checked for alignment along the microscope's optical axis. Direct alignments were performed before each image data acquisition. A list of settings used during data acquisition is presented in Table 3.13.

Micrographs were acquired semi–automatically with TVIPS$^{TM}$EM–tool which is an add–on of the image acquisition software TVIPS$^{TM}$EM–MENU (http://www.tvips.com/index.php). EM–tool uses five different lens settings (modes) for data acquisition. Before each data acquisition, each mode is checked for calibration, performed if needed, and aligned relatively to other modes.

The data acquisition pipeline is organised as follow:

- *Grid map* In this mode a map of the whole grid is acquired at magnification of 330 in LM–Mode.

**Table 3.13: FEI Tecnai$^{TM}$F20 settings.**

| | |
|---|---|
| Cathode source | FEG |
| Acceleration voltage | 200 kV |
| Gun lense | 3 |
| Extraction voltage | 3.4 kV |
| Emission | $\mu$m |
| C1 | 2 mm |
| C2 | 100 $\mu$m |
| Aperture of the objective lense | 150 $\mu$m |
| Spot size | 4 |
| Spherical aberratio ($C_s$) | 2.0 mm |
| Operation mode | low–dose |
| Exposure time | 1.5–3 s |
| Electrons/Å$^2$ | 15–30 |
| Defocus | 1–2 $\mu$m |
| **Alignments** | |
| C2 centering | ✓ |
| Objective aperture centering | ✓ |
| Gun tilt | ✓ |
| Gun shift | ✓ |
| Beam tilt | ✓ |
| Beam shift | ✓ |
| Coma-free pivot point (X, Y) | ✓ |
| Coma-free alignment (X, Y) | ✓ |
| Astigmatism | ✓ |

**Note:** Alignments were carried out after sample was loaded and eucentric height was achieved.

- *Scan* In this mode, all the positions relative to meshes suitable for data acquisition selected and acquired. Scans are acquired at magnification of 1700 in M–Mode.

- *Search* This mode is used by the software to "search" for the position, on the scan, of the area where the micrographs will be acquired. Positions are specified by user. This mode is also used by the software to correct the eucentric height. Most importantly this mode has to be setted only for the first scan since the same setting will be used for all the further scans. Search position is acquired at magnification of 5000 in SA–Mode.

- *Focus* This mode is used by the software to perform focussing. As in "search" this setting needs to be setted just for the first scan. Focussing was set at magnification of 50000 in SA–Mode.

- *Exposure* In this mode is recorded the final micrograph. As in "focus" this setting has to be specified just was since the same will be used for all the positions specified by the user. Micrograph is taken at magnification of 50000 in SA–Mode.

Once all the scans are acquired and the positions where to acquire data for each scan are specified, the software loops automatically between *search*, *focus* and *exposure* in order to record micrographs for each position.

## 3.9 Digital image processing

Image processing was supported by a blade centre with 8 nodes (64 cores) and a fast 4 terabyte file–server.

Micrographs were assessed for quality by looking at the Thon rings from the power spectrum with the aid of `e2evalimage.py` program from EMAN2, image processing software suite [Tang et al., 2007]. Micrographs with clear either drifting or charging effect or astigmatism where discarded. Eventually, only those micrographs that had the appropriate defocus, minimal astigmatism and drift were selected.

Prior boxing, micrographs were pre–processed with the program `e2proc2d.py` (EMAN2) in order to increase visual contrast and speed up picking particles. Below the image pre–processing script:

```
set t = ‘ls *.tif‘
foreach i ($h)
  e2proc2d.py $i ‘echo $i | sed -e ’s/.tif/.mrc/g’‘
    --process=filter.flattenbackground:radius=100
    --process=filter.lowpass.gauss:cutoff_freq=0.025
    --meanshrink=4
end
```

Particle picking was carried out automatically with `e2boxer.py` (EMAN2) [Tang et al., 2007]. Coordinate saved in a .box file were then used to box molecular images from the raw micrographs. Final box size was 320 x 320 pixel with a pixel size of 1.51 Å/pixel at spacemen level.

Stacks of images were created with the program `e2proc2d.py` by running a script as follow:

```
set h = ‘ls *.hed‘
foreach i ($h)
  e2proc2d.py $i stack.hed
end
```

Image processing was carried out with IMAGIC–5 [van Heel et al., 1996]. Molecular images were first normalised, band–pass filtered using a low frequency cut–off (high pass filter) 0.03 (200 Å) of and a high–pass cut–off (low pass filter) of 0.8 (7.55 Å) and then binned to 80 x 80 pixels (6.04 Å/pixel). Molecular images were then centred by translational alignment with respect to the rotationally averaged total sum of all the molecular images in the stack and iterated until the average pixel shifting was reduced almost to zero [Dube et al., 1993; van Heel et al., 2000]. The centred molecular images were subsequently aligned, using a reference-free alignment by classification protocol, in order to building a catalogue of the molecular views present in the dataset [Schatz and Van Heel, 1990]. Centered molecular images were classified using the multivariate statistical analysis (MSA) and hierarchical

classification method [Van Heel and Frank, 1981; Van Heel, 1984]. Classification can manually be influenced by choosing the number of classes to be created, usually 20–30 particles per class. It was assumed that all molecular images sorted into one class, corresponded to a certain orientation of the protein complex being analysed (ideally within a fixed conformation). Adding up all particle images within one class resulted in so-called class averages with an improved signal–to–noise ratio (SNR) [van Heel et al., 2000]. Alignment was then refined by multi reference alignment (MRA) by selecting appropriate class averages as references [van Heel et al., 2000]. Prior to the MRA the class sums were low–pass filtered and the surrounding noise was set to zero by multiplication with a soft mask of the shape of the particles. The new references were also normalized and re–centred by translational alignment with respect to their total sum. The alignment was followed another MSA leading to a new set of class averages. Refinement was repeated till no improvements were observed.

## 3.10 List of buffers and chemicals

**Table 3.14: List of buffers used in this study.**

| **Buffer** | **Ingredients** |
|---|---|
| **10X Annealing buffer** | 100 ml |
| 500 M Tris pH 8.8 | 6.05 gr |
| 500 mM NaCl | 10 ml 5M |
| Nuclease–free $H_2O$ | to 100 ml |
| pH was adjusted to 8.8 with HCl | |
| **Tris-HCl pH 6.8** | 100 ml |
| 1 M Tris | 12.1 gr |
| distilled $H_2O$ | to 100 ml |
| pH was adjusted to 6.8 with HCl | |
| **Tris-HCl pH 7.5** | 100 ml |
| 1 M Tris | 12.1 gr |
| distilled $H_2O$ | to 100 ml |
| pH was adjusted to 7.5 with HCl | |

| | |
|---|---|
| **Tris-HCl pH 8.8** | 100 ml |
| 1 M Tris | 23.8 gr |
| distilled $H_2O$ | to 100 ml |
| pH was adjusted to 8.8 with HCl | |
| **HEPES-KOH pH 7.4** | 100 ml |
| 1 M HEPES | 12.1 gr |
| distilled $H_2O$ | to 100 ml |
| pH was adjusted to 7.4 with KOH | |
| **HEPES-KOH pH 8.0** | 100 ml |
| 1 M HEPES | 12.1 gr |
| distilled $H_2O$ | to 100 ml |
| pH was adjusted to 8.0 with KOH | |
| **Magnesium Chloride ($MgCl_2$)** | 50 ml |
| 1 M $MgCl_2$ | 4.76 gr |
| distilled $H_2O$ | to 50 ml |
| **MES-HCl pH 6.0** | 100 ml |
| 1 M MES | 19.5 gr |
| distilled $H_2O$ | to 100 ml |
| pH was adjusted to 6.0 with KOH | |
| **5M NaCl[1]** | 1 L |
| 5 M NaCl | 58.44 gr |
| distilled $H_2O$ | to 1 L |

**Table 3.15: List of chemicals and ragents used in this study.**

| Abbr | Chemical | Company | Purity |
|---|---|---|---|
| Acrylamide | 30% Acrylamide bis solution (37.5:1) | Severn Biothec | EP |
| UltraPure™ | Agarose | Invitrogen | EP |
| | Acetic acid | Fisher | ARG |
| APS | Ammonium persulfate | Sigma | ≥98% |
| | Bacto-tryptone | Bacto | |
| | Yeast extract | Oxoid | |
| BFB | Bromophenol blue | Sigma | EP |
| | 100bp DNA ladder | NEB | |
| | 1kb DNA ladder | NEB | |
| DMSO | Dimethylsulfoxide | Sigma | 99% |
| EtOH | Ethanol | UniEd stores | |
| EtBr | Ethidium bromide solution (10 mg/ml) | Sigma | ≥95% |
| EDTA | Ethylendiamintetra-acetic acid | Sigma | ≥98% |
| | Formaldehyde 37% | Sigma | ACS |
| | Glutaraldehyde (25%) | Sigma | 99% |
| | Glyerol | Fisher | LRG |
| | Glycine | Sigma | >99% |
| HEPES | Hydroxyethyl-monosodium salt | Sigma | ≥99.5% |
| HRP-Ab | ECL Rabbit IgG, HRP-linked whole Ab (from donkey) | GE | |
| | Imidazole | Acros | >99% |
| $MgCl_2$ | Magnesium chloride anydrous | Sigma | ≥98% |
| MeOH | Methanol | Fisher | ARG |
| Anti-His | Monoclonal Anti-polyHistidine | Sigma | |
| TEMED | N,N,N',N'-Tetramethylethylene-diamine | Sigma | >99% |
| NF | Nulease-free water | Sigma | >99% |
| PIPES | Piperazine–N,N'–bis(2–ethanesulfonic) acid | Sigma | >99% |
| PMSF | Phenylmethannesulfoniylfluriode | Sigma | >99% |
| Tween20 | Polyoxyethylene sorbitan monolaurate | Sigma | |
| $KH_2PO_4$ | Potassium phosphate monobasic | Sigma | >99% |
| $K_2HPO_4$ | Potassium phosphate dibasic | Sigma | >99% |
| PIC | Protease inhibitor EDTA-free | Roche | |
| $AgNO^3$ | Silver nitrate | Sigma | >99% |
| | SuperSignal Western Blotting Kits | Pirce | |
| BCIP®/NBT | 5-Bromo-4-chloro-3-indolyl phosphate/ Nitro blue tetrazolium | Sigma | |
| | Skim milk | Oxoid | |
| NaCl | Sodium chloride | Fisher | ARG |
| SDS | Sodium dodecylsulfate 10% | Fisher | EP |
| | $\beta$-mercaptoethanol | Sigma | 99% |
| UltraPure Tris | Tris-(hydroxymethyl)-aminomethane | Invitrogen | >99% |
| UA | Uranyl acetate | Fisons | ARG |

EP. elctrophoresis purity. LRG, laboratory reagent grade. ARG, analytical reagent grade. ACS, American Chemical Society reagent.

# Chapter 4

# Results and Discussion: sequence analysis, cloning, purification and binding assay of *Pab*MCM helicase

## 4.1 Introduction

Several archaeal genomes have been sequenced to date. Bioinformatics analysis suggest a degree of homology between eukaryotic replication, trascription and translation proteins [Edgell and Doolittle, 1997]. Thus, Archaea provide us with a simplified model for understanding complex molecular machinery involved in DNA metabolism [Barry and Bell, 2006].

Homologues of eukaryotic MCM protein complex have been identified in all sequenced archaeal genomes, which have one MCM–like protein [Jenkinson and Chong, 2003]. Exceptions are *Methanococcus jannashii*, which posses four MCM protein complexes while *Methanosarcina acetivorans* and *Methanococcus kandleri* have two MCM–like proteins [Bult et al., 1996; Galagan et al., 2002; Slesarev et al., 2002].

The *P. abyssi* genomic ORF PAB2373 was investigated, which has been predicted to encode a MCM–like protein.

## 4.2 Sequence analysis of the *Pab*MCM's AAA$^+$ catalytic domain

The sequence alignment of the archaeal MCM proteins from *S. solfataricus*(Q9UXG1),*M. Thermoautrophicus*(O27798) and *P. furiosis*(Q8U3I4) with the predicted one from *P. abyssi* reveals high sequence homology (Figure 4.1).

The sequence alignment also shows the presence of two insertions. Based on sequence analysis, these sequences were predicted to be inteins. Inteins, also called protein introns, are proteins capable of self–excising from the host protein (the exteins), in a process called protein splicing [Gogarten et al., 2002]. *Pab*MCM has two intein domains. The first *Pab*MCM intein is inserted into the C–terminal of the Walker A motif, whereas the second one in inserted at the N–terminus H2I–hp motif. Interestingly, *Sso*MCM and *Mth*MCM has no intein domains while in *Pfu*MCM has just one inserted at the N–terminus H2I–hp motif.

The N–terminus domain of the *Pab*MCM reveals low sequence similarity with previously reported *S. solfataricus* and *M. Thermoautrophicus* MCM proteins (Figure 4.1). The N–terminus has been suggested to possess regulative roles as mutational studies have shown previously [Kasiviswanathan et al., 2004]. The N–terminus domain is involved in the formation of the single ring as well as for the double ring of MCM complex [Fletcher et al., 2003; Chong et al., 2000]. Fletcher et al. [2003] revealed the three–domain (A, B and C) structure of the N–terminus domain. Biochemical and biophysical characterisations of the N–terminus domain have shown that it plays a regulatory role in MCM function [Barry et al., 2009; Sakakibara et al., 2008; Kasiviswanathan et al., 2004]. Speculating on this observation, the N–terminus domain of *Pab*MCM might have a regulatory role as well. The little sequence similarity of the N–terminus of the A and B domains could represent a form of adaptation to the extreme environment while the higher conservation of the C domain and the NCL linker a similar role in the ring formation and intersubunits communication as previously seen for other MCM protein complexes.

The MCM protein complex belongs to the AAA$^+$ family of ATPase [Ogura and Wilkinson, 2001a]. The AAA$^+$ catalytic domain of *Pab*MCM is better conserved

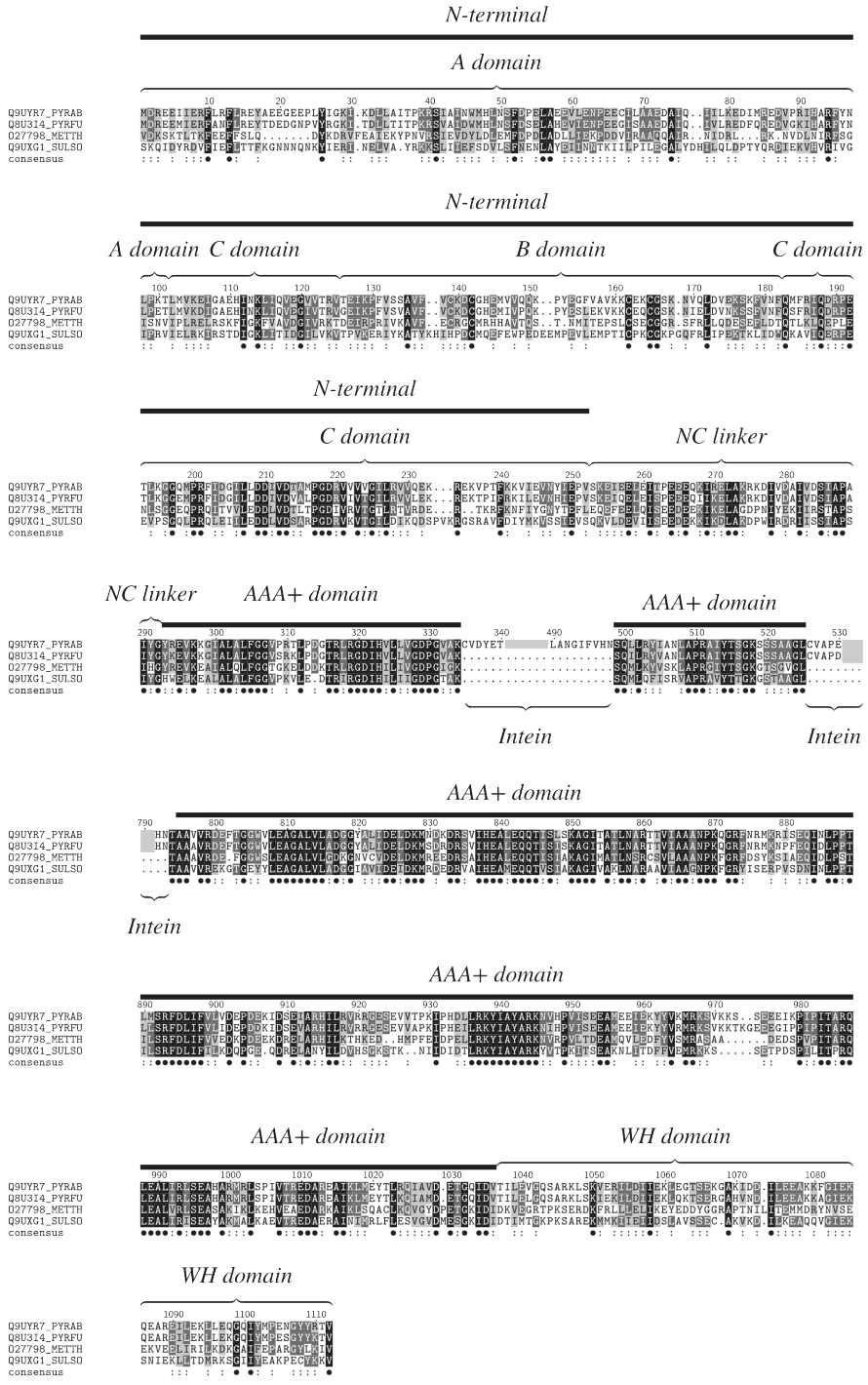



**Figure 4.1: Sequence alignment of *S. solfataricus*(Q9UXG1),*M. Thermoautrophicus*(O27798), *P. furiosis*(Q8U3I4) and *P. abyssi* MCMs.** Highly conserved residues (indentity 100%) are shaded in black while black dots are shown in the consensus. Residues with indentity >75% are shaded in light grey while colons are shown.

compared to the N–terminus (Figure 4.2). As reported in Barry et al. [2007] and Grainge et al. [2006] this domain forms hexamers and possesses helicase activity. In the *Pab*MCM ATPase, the active site ($AAA^+$ domain) is ∼250 residues long. Important motifs, which impart catalytic activity, are Walker A, Walker B, sensor 1 (S1) and sensor 2 (S2). The Walker A motif is involved in binding ATP Walker B and S1 orient the nucleophilic water molecule. S2 with the arginine–finger motif contact the $\gamma$–phosphate [Bochman and Schwacha, 2009]. The sequence alignment shows that all these active–site motifs are well conserved in *Pab*MCM although polymorphisms are present in functional motifs such as EXT–hp, Walker A, H2I–hp, PS1–hp. These differences could represent mutations evolved as a form of adaptation to the different extreme habits in which they thrive. This strong sequence homology with the two well–known archaeal MCM proteins, strengthens the possibility that the ORF PAB2373 is the replicative helicase of *P. abyssi*.

### Codon–usage analysis

The genetic code is degenerate Kornberg and Baker [1992]. Accurate and efficient translation of highly expressed genes is a result of codon–usage bias, which has been observed in almost all genomes [Ikemura, 1985; Plotkin and Kudla, 2010]. Optimal heterologous gene expression requires a *priori* knowledge of the host organism codon–usage.

The codon usage analysis of *Pab*MCM was performed using the online web–server Graphical Codon Usage Analyser (http://gcua.schoedl.de/). The analysis was performed against *E. coli* codon–usage table (Figure 4.3) as this was the first host cell choice.

As results show in Figure 4.3, *Pab*MCM has a codon–usage slightly different (mean difference ∼12%) compared to *E. coli* codon usage. In fact, comparing the codons for the amino–acid arginine, it is quite clear that *E. coli* will use more efficiently CGT and CGC codons rather than AGA and AGG ones. In terms of efficiency of translation this can be translated in lower level of protein expression. To increase *Pab*MCM protein expression, Rosetta$^{TM}$(DE3) pLysS cells were tested first as host strain for protein expression. Rosetta$^{TM}$(DE3) pLysS host strain are

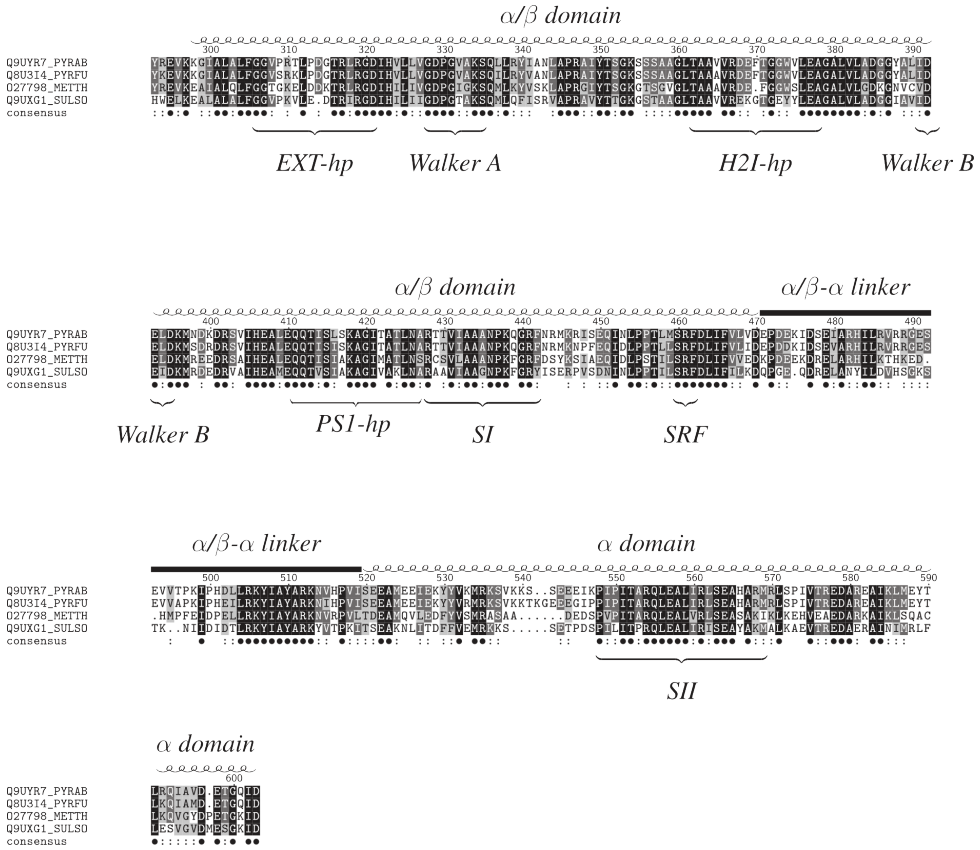



**Figure 4.2: Structural conservation of *Pab*MCM AAA$^+$ catalytic domain.** *S. solfataricus*(Q9UXG1),*M. Thermoautrophicus*(O27798), *P. furiosis*(Q8U3I4) and *P. abyssi* MCMs were aligned. Highly conserved residues (indentity $\geq$90%) are shaded in black while black dots are shown in the consensus. Residues with indentity $\geq$70% are shaded in light grey while colons are shown.

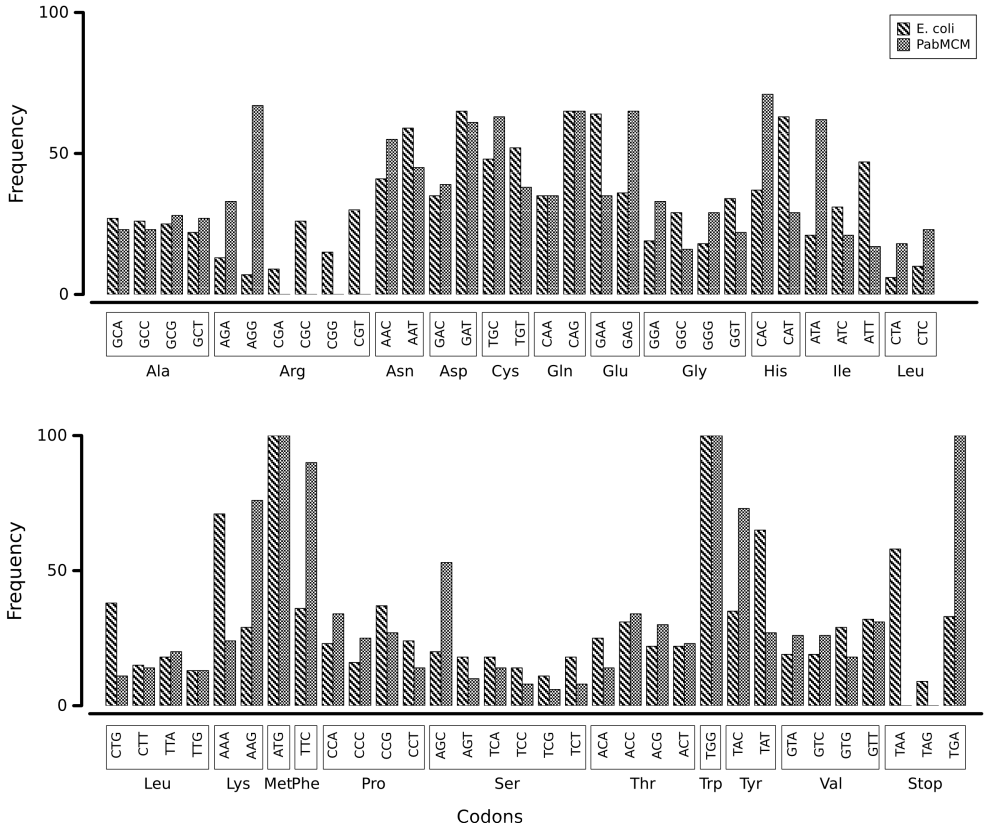



**Figure 4.3: Codon usage analysis of the *Pab*MCM ORF.** $Pab$MCM full-length codon–usage was compared against *e. coli* codon–usage table in order to detect whether $Pab$MCM full-length had different codon–usage and hence affect the translation rate during protein expression.

designed to enhance the expression of proteins that contain codons rarely used in *E. coli*. These strains supply tRNAs for AGG, AGA, AUA, CUA, CCC, GGA codons on a compatible chloramphenicol–resistant plasmid.

## 4.3 Cloning of the archeal *Pab*MCM gene

The ORF PAB2373 (3336 bp) deposited at the NCBI (http://www.ncbi.nlm.nih.gov/) data bank, encodes a protein of 1112 amino acid residue and it has been predicted to be one of *Pab*MCM proteins. As mentioned, the ORF contains two intein domains (Figure 4.1), whose insertion in the sequences cause the isolation of a fragment of the 27 amino acid residues. As result of this insertion, the N–terminus Ser–499 is isolated from the adjacent Lys–525 of the predicted Walker A domain while the N–terminus sequence of the predicted H2I–hp is not being affected (Figure 4.1).

Accordingly to the sequence alignment (Figure 4.2), the putative active full–length *Pab*MCM is devoid of inteins. In order to reconstruct the full–length gene, a PCR–based approach, in which fragments of the putative coding sequence were amplified separately and fused by PCR, was used (Figure 4.4).

The pair of primers F_PAB2373–1/R_PAB2373_in was used to PCR–amplify the N–terminus of MCM (Figure 4.5A), the pair of primers J–NTERM and J–CTERM for amplify the fragment Ser499–Leu525 (Figure 4.5B) and finally the pair of primers F_PAB2373_in and R_PAB2373 was used for PCR–amplify the C–terminus fragment of MCM (Figure 4.5A).

The full–length *Pab*MCM was reconstructed in three PCR runs (Figure 4.4(II)). In each run, the PCR to join the fragments was performed in two steps. In the first step, two fragments were incubated for 5 cycles, in absence of primers, in order to fill the 5' ends. In the second step, a mix containing primers was added and PCR was performed for 35 cycles. In the first run, the fragment N–terminal was fused with the fragment Ser499–Leu525 and then amplified by F_PAB2373–1/C–TERM (Figure 4.5C). In the second run, the fragment Ser499–Leu525 was fused with the C–terminal fragment of MCM and then amplified by N–TERM/R_PAB2373 (Fig-

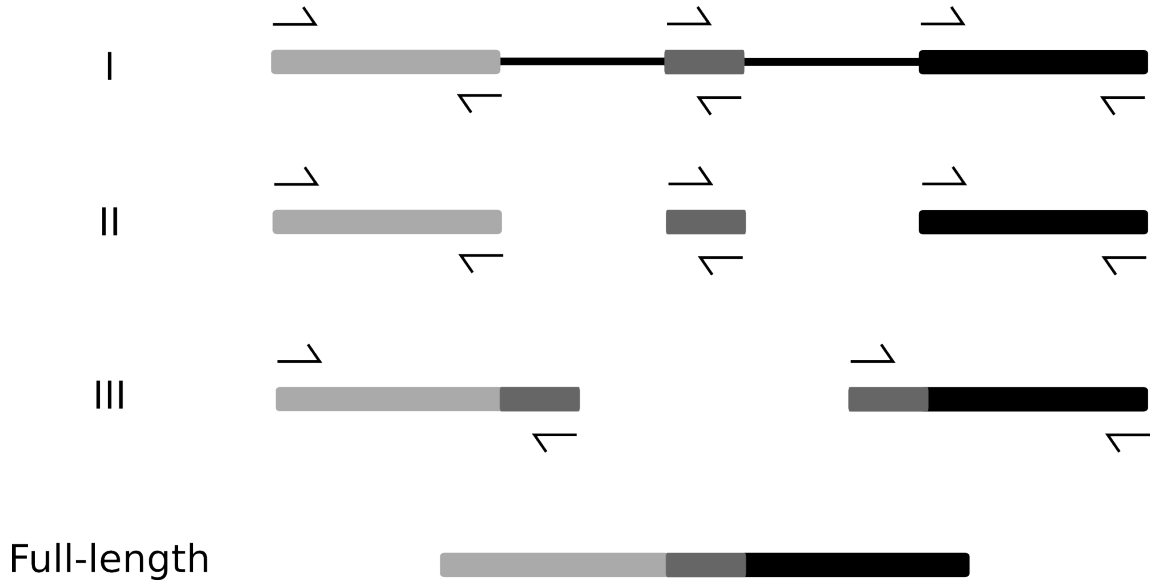



**Figure 4.4: Schematic representation of PCR-based strategy used for reconstructing the full-legnth *Pab*MCM.** (I) Schematic of the ORF2373 from *P. abyssi*. (II) 5'-end, internal and 3'-end fragments amplified separately. (III) 5'-end joined to internal fragment and internal fragment joined to 3'-end. (Full-length) Final PCR to join both fragment from III.

ure 4.5D). Finally, in the third run the fragment N–Terminal–Ser499–Leu525 and Ser499–Leu525–C–terminal were fused by F_PAB2373/R_PAB2373 (Figure 4.5E). To confirm that the fragment had been inserted in the sequence, a PCR was performed by using the couple of primers J–NTERM and J–CTERM (Figure 4.5F) .

The reconstructed full–length MCM fragment was cloned into the pET3a–Tr vector. The *P. abyssi* full–length MCM ORF was confirmed by sequencing (Figure 4.6). In the sequence, neither frame shift nor mutations were found.

## 4.4 Expression and purification of the recombinant archeal *Pab*MCM

### 4.4.1 Preliminary expression screenings

**over–expression test** Small scale over–expression tests were carried out in experiments using BL–21 CodonPlus$^{TM}$(DE3)–RIPL transformed with the construct pET3a–tr–MCMFL.

The result seen in Figure 4.7, shows that the *Pab*MCM protein is expressed (lines 1,2,3,4,16). The calculated MW is 74.2 kDa and the apparent MW in SDS–PAGE is in agreement with this. The *Pab*MCM protein seems to undergo a proteolytic cleavage since a smaller band, running around 60 kDa, is seen in the gel. No detectable changes in the expression of the *Pab*MCM, were seen after 1 hour expression throughout the time of induction. Basal protein expression is detectable (line NI). For this reason, the cell strain Rosetta$^{TM}$(DE3) pLysS was used in further experiments. This strain harbours the vector pLysS which encodes T7 lysozyme, which allows a tight control on the T7 RNA polymerase. T7 lysozyme is a natural inhibitor of T7 RNA polymerase, and thus reduces its ability to transcribe target genes in uninduced cells.

**Heat–denaturation test** *P. abyssi* is a hyperthermophilic archaeon, whose optimal habit for thriving is the deap–sea hydrothermal vent [Cohen et al., 2003]. To

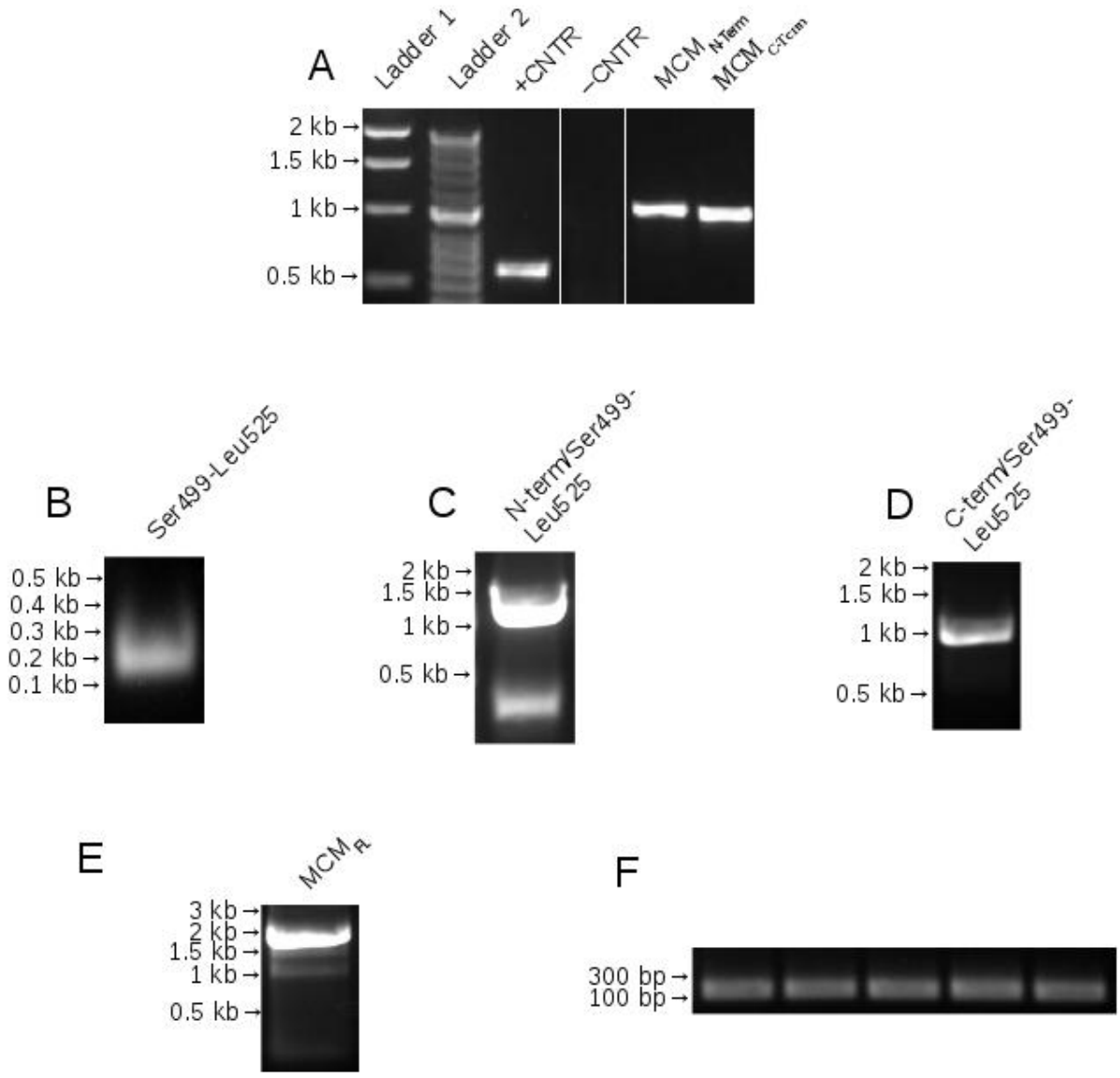



**Figure 4.5: Recostruction of the full-length *Pab*MCM devoided of inteins**. (A) PCR-amplified N and C-terminus. (B) PCR-amplified fragment Ser499-Leu525. (C) PCR-amplified of the joined N-term/Ser499-Leu525. (D) PCR-amplified of the joined N-term/Ser499-Leu525. (E) PCR-amplification of the $Pab$MCM full-length. (D) PCR performed in order to check that the fragment was successfully joined into the MCM sequence.

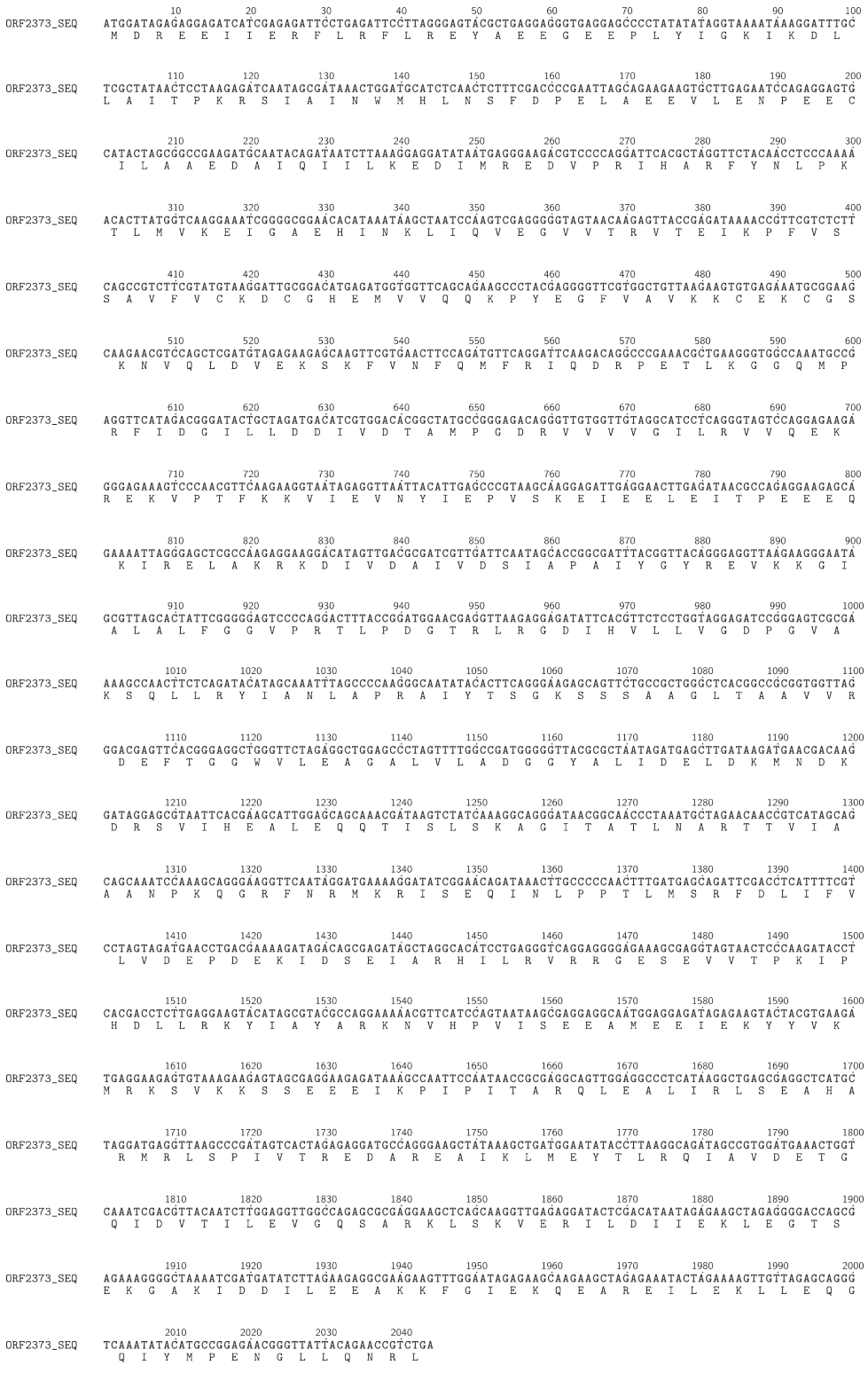



**Figure 4.6: Sequencing result for the reconstructed full-length *P. abyssi* MCMs (ORF PAB2373).** The vector pET3a–Tr harbouring the reconstructed MCMFL was sequenced at the GenePool sequencing facility at Edinburgh University. 6 primers were used for generating forward and reverse reads of 500 bp each in order to have good sequence coverage. Here the final conting is being shown. The contig was generated with CAP3 Sequence Assembly Program (http://doua.prabi.fr/software/cap3)

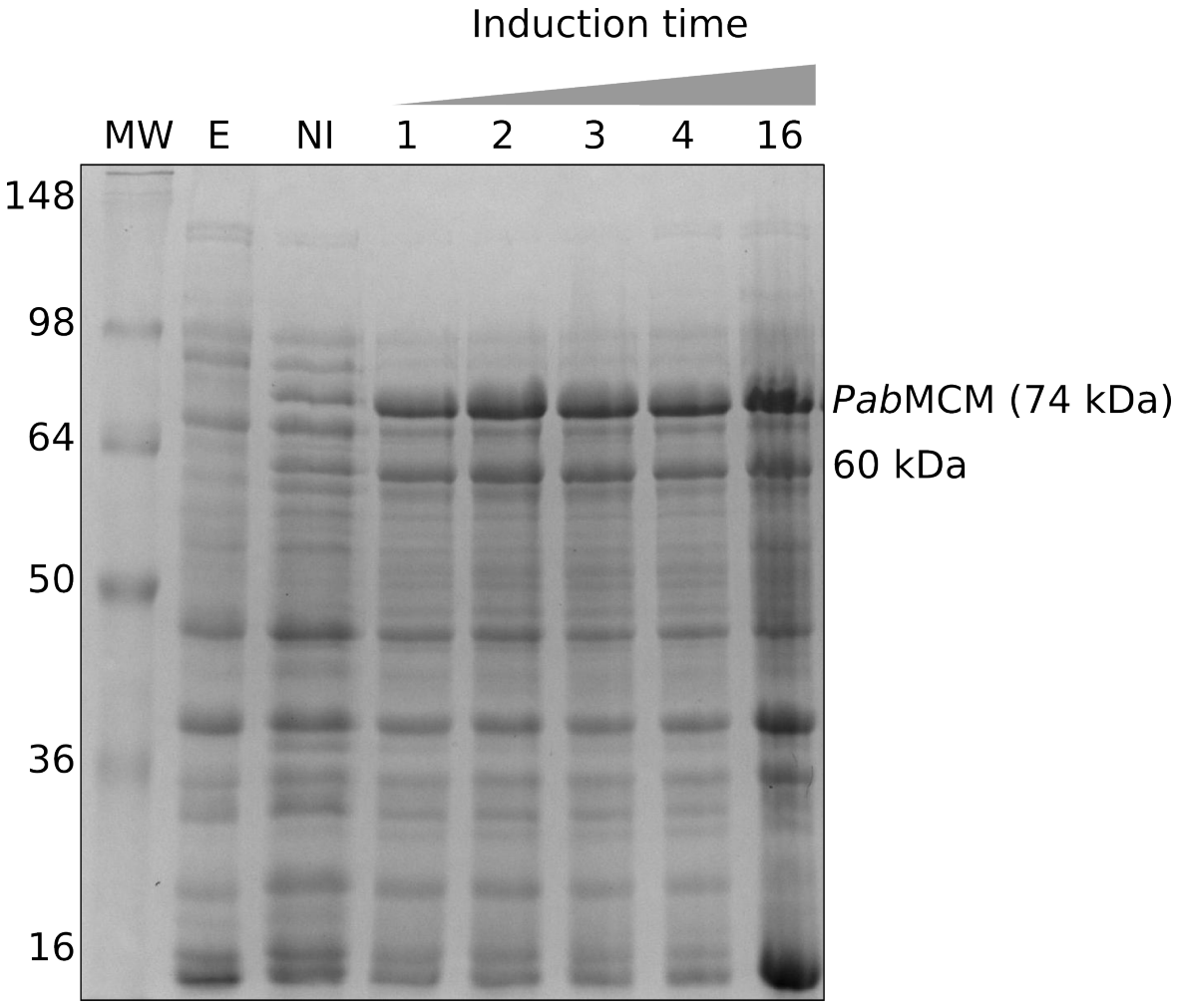



**Figure 4.7: Small scale over-expression test of *Pab*MCM.** (MW) Protein ladder. (E) BL-21 CodonPlus$^{TM}$(DE3)-RIPL without construct. (NI) Not induced BL-21 CodonPlus$^{TM}$(DE3)-RIPL. (1,2,3,4,16) Post-induction harvesting time expressed in hours.

test *Pab*MCM thermostability, cell lysate was incubated at different temperatures (Figure 4.8). As shown, the largest amount of the expressed *Pab*MCM protein is in the pellet (line RT(P)). This suggests partial insolubility, which is most likely due to the high–level of recombinant protein expression resulting in accumulation of insoluble aggregates in inclusion bodies [Kane and Hartley, 1988]. However, a good fraction of *Pab*MCM is still soluble at room temperature (line RT(SN)). In this experiment, the high thermostability of *Pab*MCM it is also shown (Figure 4.8). After heating the whole lysate (up to 90°C) the *Pab*MCM protein is still soluble. Little changes are visually detected in the band intensity compared with the soluble fraction not heat–treated. The heating does not seem to have any effect on the band running around 60 kDa. Importantly, this test shows how heat–denaturation is an important purification step since >90% of the host proteins become insoluble after heat treatment. Heat–denaturation could be considered as a proper purification step, when working with hyperthermophilic organisms, since host proteins, which are not thermostable will denature and hence separate by a simple centrifugation step.

**Chromatography: heparin and ion–exchange binding tests** Previous studies on archaeal MCM protein complex have reported a binding affinity to both heparin and strong cation exchange resins [Yoshimochi et al., 2008]. Moreover, it is already known that DNA binding proteins quite often have affinity for heparin resin [Xiong et al., 2008]. Small scale binding tests were performed in order to test and choose the right chromatography strategy for purification of *Pab*MCM (Figure 4.9). Heparin, strong cation exchange and strong anion exchange slurry resin were used for this experiment. In Figure 4.9, the eluted fractions are shown. As expected with DNA–binding proteins, *Pab*MCM binds heparin (line H) as well as the strong cation exchange resin (line Q). No affinity binding was seen when the protein was incubated with a strong anion exchange resin (line S).

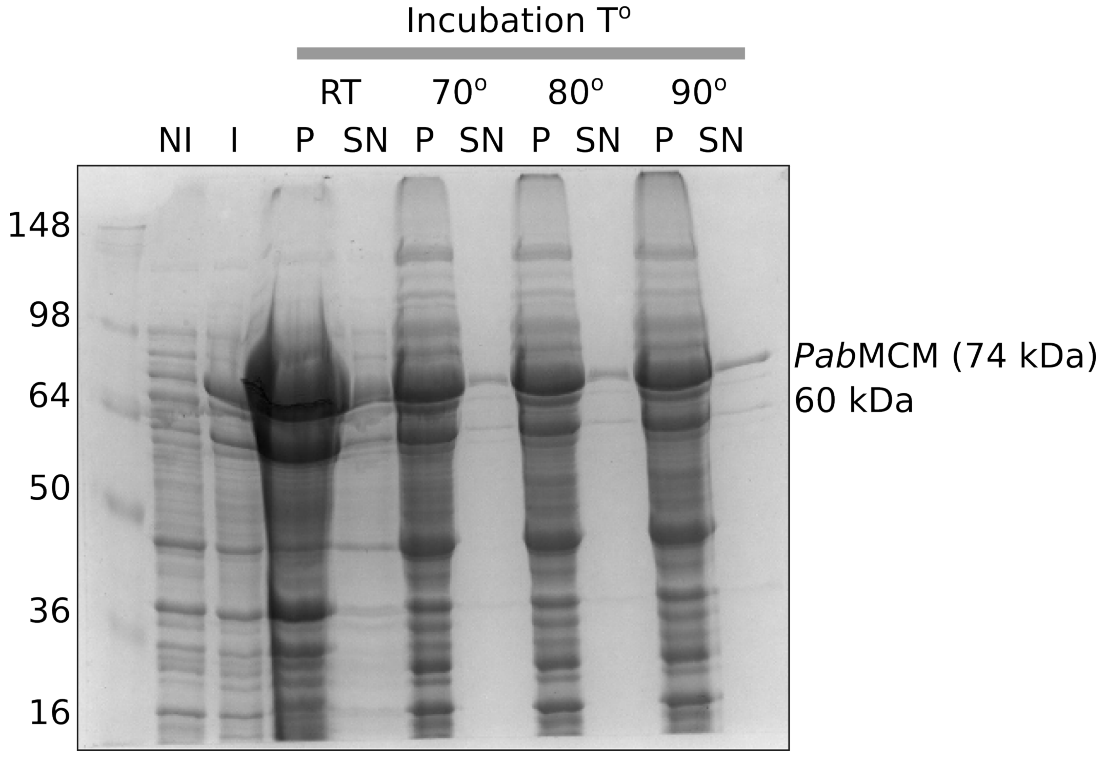



**Figure 4.8: Heat-denaturation test for *Pab*MCM.** (MW) Protein ladder. (NI) Not induced BL-21 CodonPlus$^{TM}$(DE3)-RIPL. (I) Induced BL-21 CodonPlus$^{TM}$(DE3)-RIPL. (P) Cell pellet after cell lysys and centrifugation at 13000 rpm. (SN) Supernatant obtained as in (P). (RT, 70°C,80°C,90°C) Temperature at which cell lysate was incubated in order to test the *Pab*MCM heat-resistance; RT indicates room temperature.

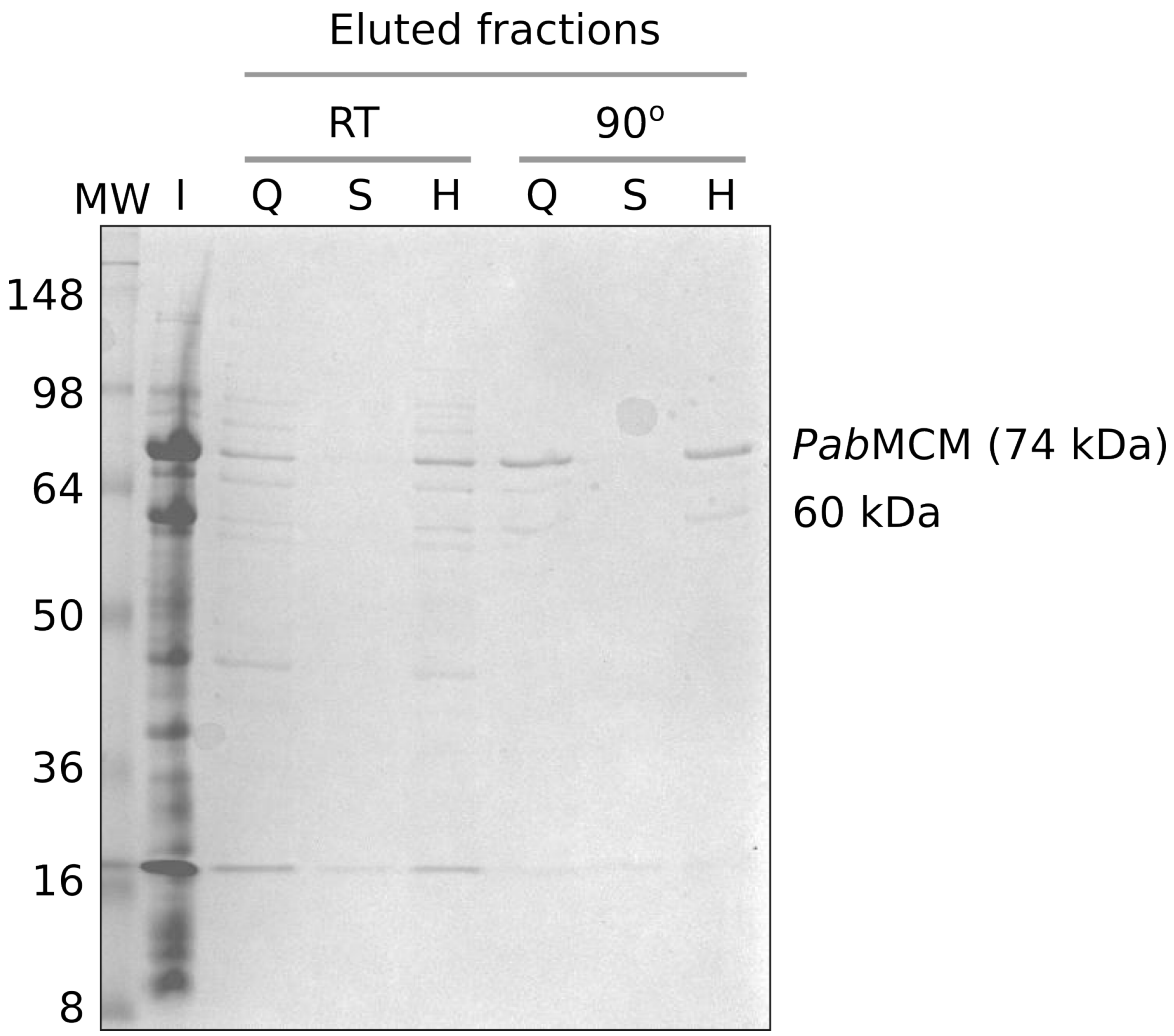



**Figure 4.9: Chromatography binding test of *Pab*MCM.** Soluble *Pab*MCM not heat-treated (RT) and heat-treated (90°C) fractions were tested for binding capacity to three types of resins: heparin (H), strong cation exchange (S) and strong anion exchange (Q). (MW) Protein ladder. (I) Input protein.

### 4.4.2 Protein purification protocol optimisation

Many strategies were tested in order to separate the 60 kDa co–purified fragment from the protein preparation. Several chromatographic techniques were tested and optimised (hydrophobic interaction chromatography, weak ion-exchange chromatography, ion-exchange chromatography) for that purpose (data not shown). Five of these strategies will be described.

Figure 4.10 (A) shows five size–exclusion chromatography traces related to five attempts carried out in study.

The first experiment (Native in panel A and B(I)), is representative of the very first purification protocol. The sample, after elution first from a heparin affinity column and from a strong cation exchange (data not shown), was loaded into a size–exclusion chromatography column. The sample eluted nearly as single peak although a smaller peak came off as well. Collected fractions were analysed on SDS–PAGE and shown that the sample was almost pure. However the smaller 60 kDa protein was co–eluting with *pab*MCM. Eluted fractions were also checked at the microscope to test whether or not the purified protein was forming any ring–shape assembly. As shown (panel B(I) from F11 to F12) a ring–shape like assembly could be seen at the microscope. As shown in Figure 4.8, a fraction of the expressed *Pab*MCM is in the inclusion bodies. Protein purification from inclusion bodies has been successfully applied in number of recombinant proteins [Vallejo and Rinas, 2004]. In order to test whether or not this fraction could be recovered by refolding, Inclusion bodies were resuspended in 6 M GuCl and dialysed against 20 mM HEPES pH 7.4, 500 mM NaCl overnight at room temperature. As shown in Figure 4.10 (panel B IncBodies II), *Pab*MCM could be solubilised and refolding was successful as demonstrated in the related micrographs.

Affinity–tag purification was another option taken into account in order to separate the small 60 kDa co–eluting protein. *Pab*MCM reconstructed full–length ORF was sub–cloned into a modified pET vector with N–terminus six–histidine tag and a TEV protease cleavage site (pETWO–E). The new construct was sequenced and tested to be effectively functional and proteins expression was checked in small–scale protein expression tests and small scale nickel affinity chromatography were

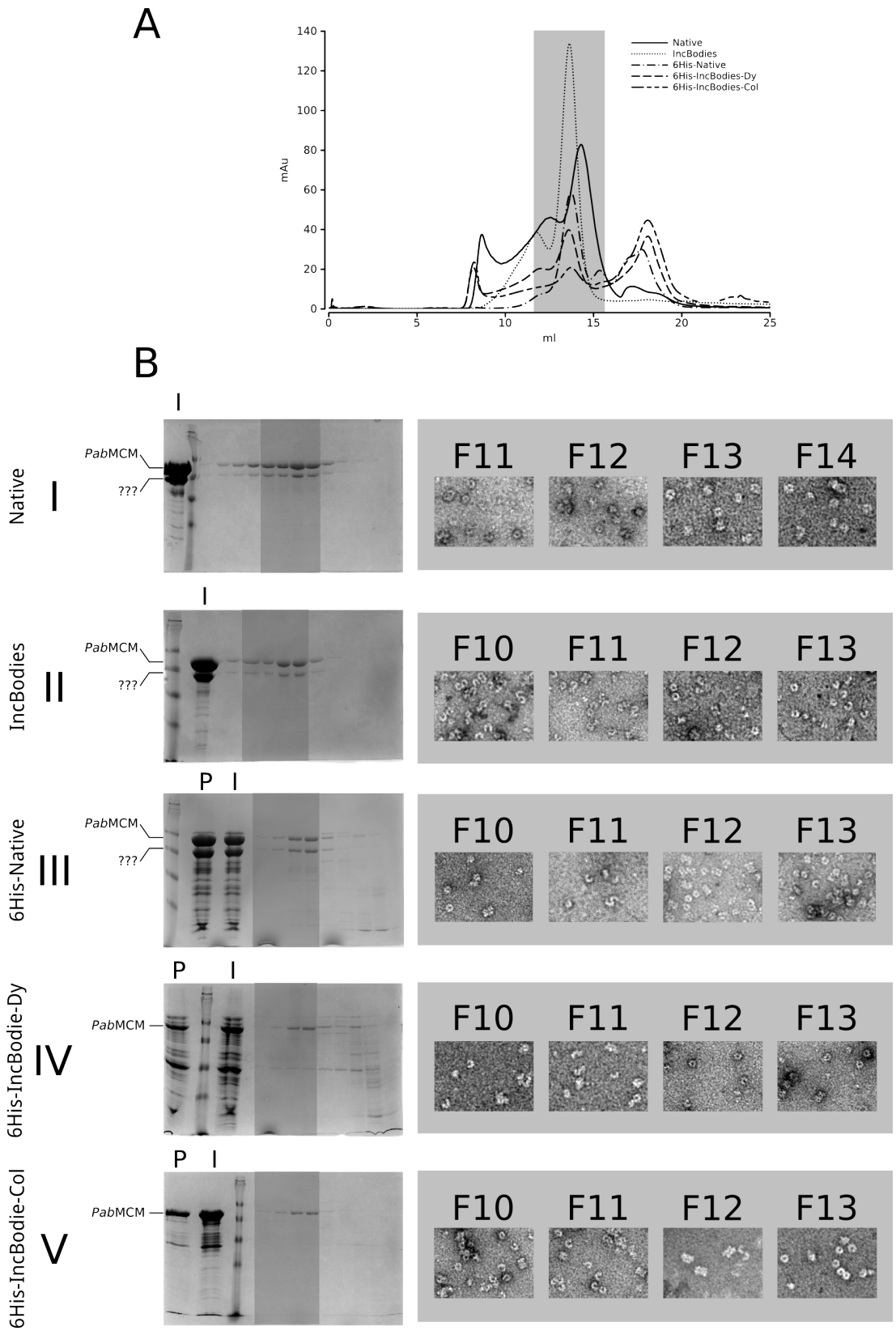



**Figure 4.10: Purification protocols scouting for *Pab*MCM.** (A) Overlapped size-exclusion chromatography traces of *Pab*MCM carried out in 20 mM HEPES pH 7.4, 150 mM NaCl at 0.3 ml/min flow-rate onto a Superose6$^{TM}$10/300 GL. (B) SDS–PAGE and electron microscopy analysis of *Pab*MCM fractions.

performed in order to test the binding (data not shown). In Figure 4.10 (6His-Native in panel A and B III), it is shown the size–exclusion chromatography carried out with the 6xHis–tagged *Pab*MCM protein. As shown, affinity purification did not help in separating the 60 kDa fragment since still two band appeared when eluted fractions were analysed on SDS–PAGE. Electron microscopy analysis showed ring–shape like particles. Yet, few more elongated particles, which could resemble the double ring assembly previously seen for other archaeal MCM–like protein complex [Costa et al., 2006a; Gómez-Llorente et al., 2005], were also seen in the microscope . Denaturing conditions were tested with the 6–His tagged *Pab*MCM. The rational behind this experiments was to test whether or not the full–length *Pab*MCM could be separated from the 60 kDa protein by using harsh denaturing conditions. As shown above (Figure 4.10 (panel B IncBodies II)), *Pab*MCM could be refolded in a soluble ring–shaped protein complex. In this case two strategies were adopted. In the first approach (Figure 4.10 panel A and B IV, 6His–IncBodies–Dy), the resolubilised 6–his tagged *Pab*MCM was refolded by dialysis at room temperature overnight. The refolded protein adopted a ring–shape like assembly after refolding, meaning that *Pab*MCM was most likely folding nearly to a native state (Figure 4.10 F10–F13). In the second approach (Figure 4.10 panel A and B V, 6His–IncBodies–Col), the 6–his tagged *Pab*MCM was resolubilised from inclusion bodies and refolded on–column by gradual buffer exchange to 20 mM HEPES pH 7.4, 500 mM NaCl.

From these experiments, it is possible to conclude that the best way to purify *Pab*MCM without the small 60 kDa fragment is to use an affinity tag chromatography in denaturing conditions. In this specific study, the N-terminal 6–histidines tag could be used since its capacity to retain binding to a nickel column even in denaturing conditions. Since the small 60 kDa fragment is lacking of the N-terminal (see 4.4.3), it will not bind the to the nickel column. Yet, the denaturing conditions will help to dissassemble the ring in single monomeric units and hence the 60 kDa fragments will elute with the flow–through. Moreover, these data suggest that the best way to obtained a pure refolded *Pab*MCM is to perform refolding in column.

### 4.4.3 Mass–spectrometry analysis confirms that both bands are *Pab*MCM

Protein–fingerprinting, which is a method for protein identification via mass–spectrometry [Pappin et al., 1993; Mann et al., 1993], was adopted to confirm that the band running around 70 kDa was *Pab*MCM as well as to identify the 60 kDa co–purified product (Figure 4.7). Results of this experiment are shown in (Figure 4.11). This analysis confirmed, that the higher MW band was *Pab*MCM while the 60 kDa band, was a shorter form of *pab*MCM. No peptides were detected for the N–terminus of the 60 kDa band. This may mean that the 60 kDa *Pab*MCM was probably missing the N–terminus.

### 4.4.4 Prediction and Mutagenesis of a putative intragenic MCM's Shine–Dalgarno sequence

Preliminary expression tests in Rosetta$^{TM}$(DE3) pLysS harbouring the plasmid vector pET3–tr–MCMFL shown a secondary shorter protein product being co–expressed with the full–length *Pab*MCM (Figure 4.7).

Ribosome–mediated protein synthesis is regulated at the level of initiation of translation. Different mechanisms of initiation of translation and regulation of protein expression levels are adopted between prokaryotes and eukaryotes. In prokaryotic mRNA, initiation of translation is controlled via structural elements as Shine–Dalgarno sequence nearby the AUG and AUG itself. [Matten et al., 1998; Kozak, 2005; Tikole and Sankararamakrishnan, 2006]. The Shine–Dalgarno (SD) sequence is a short 5'–ACCUCC–3' motif of 3' end of *E. coli*'s 16S ribosomial RNA (rRNA), which is known to be complementary to the 5'–GGAGGU–3' motif located at the 5' end of several messenger RNAs (mRNA). This short motif has been shown to be sufficient to create a stable double stranded nucleic acid structure that could position the ribosome correctly on the mRNA during translation initiation [Shine and Dalgarno, 1974]. Mass–spectrometry analysis shown that the shorter protein running around 60 kDa was a truncated form of *Pab*MCM full–length. Since there were no indications of either degradation or proteolysis and in light of what has been

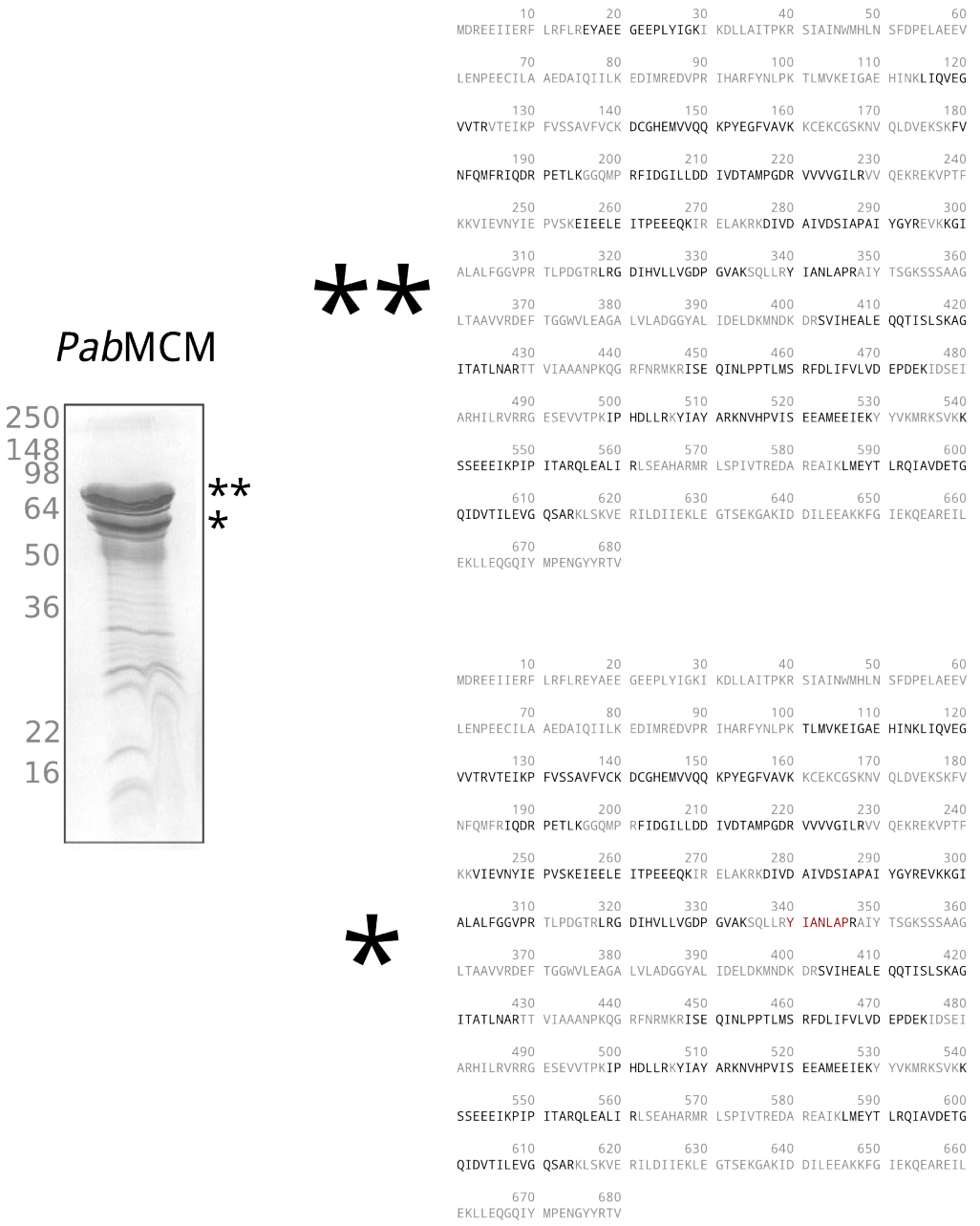



**Figure 4.11: Mass-spectrometry analysis of *Pab*MCMs purified proteins.** (A) SDS–PAGE of *Pab*MCM, ⋆⋆ full-length *Pab*MCM while in (B) it is shown the peptide coverage peptide (colored grey). As shown, no pepties peptides, belonging to the N–terminal, were seen to fly in the MALDI–TOF when the shorter form was analysed while the coverage for the C-terminal remained the same. Importantly, the peptide YIANLAP (in red), which is part of the $AAA^+$ domain being isolated by the two inteins insertions in the *Pab*MCM sequence (Figure 4.2), was detected. This observation clearly indicates that the full-length protein was reconstructed properly.

reported above, alternative intragenic start codons as well as alternative ribosome binding sequences (RSD) were investigated.

Based on the difference in molecular weight between the *PabMCM* (74 kDa) and the shorter polypeptide of *Pab*MCM (∼60 kDa), the first 5' 306 nt of the *Pab*MCM ORF were analysed. Since it is not possible to accurately estimate the molecular weight on SDS–PAGE, a bigger range from the first methionine to the fourth one (∼25 kDa encoded peptide) was taken into account. Bioinformatics analysis performed with the software Free2Bind [Starmer, 2000] was used to screen the first 5' 306 nt of *Pab*MCM. Free2Bind found three putative RBSs in the sequence (Figure 4.12).

The first match ($\Delta G = -11.22$) was too close to the first start codon also no AUG were seen in the 17 nt downstream, which represent the optimal distance of the SD from the AUG as shown in Chen et al. [1994].

The second match ($\Delta G = -10.14$) revealed an SD–like sequence 5 nt upstream of the third in–frame AUG codon. This sequence is identical to initiation of translation elements found in the pET system vectors on the market. This is most likely the reason why the shorter *Pab*MCM was being co–expressed. To validate this hypothesis two experiments were carried out. In the first experiment, *Pab*MCM was sub–cloned in the vector pETM–40, with an N–terminus maltose binding protein and a C–terminal 6–His tag. Shifting downstream this alternative initiation of translation element would have greatly reduced the expression of the shorter fragment. As shown in Figure 4.13B, no shorter fragment could be seen in SDS–PAGE. This result was also confirmed via western blotting by probing the membrane with an anti–His tag antibody Figure 4.13C. The western blotting also reveal that neither degradation nor proteolysis is occurring since no smeared bands were seen. However, this construct could not be used for further experiments since further analysis revealed that the MBP tag was not cleavable. In addition, heat–denaturation would not be applicable. The second experiment performed was the mutagenesis of the 5'–AAGGAGGATATA–3' sequence to 5'–AAAGAAGACATT–3'. As shown in Figure 4.12B, this new construct was being expressed and no co–expressed products were seen when samples were analysed on SDS–PAGE.

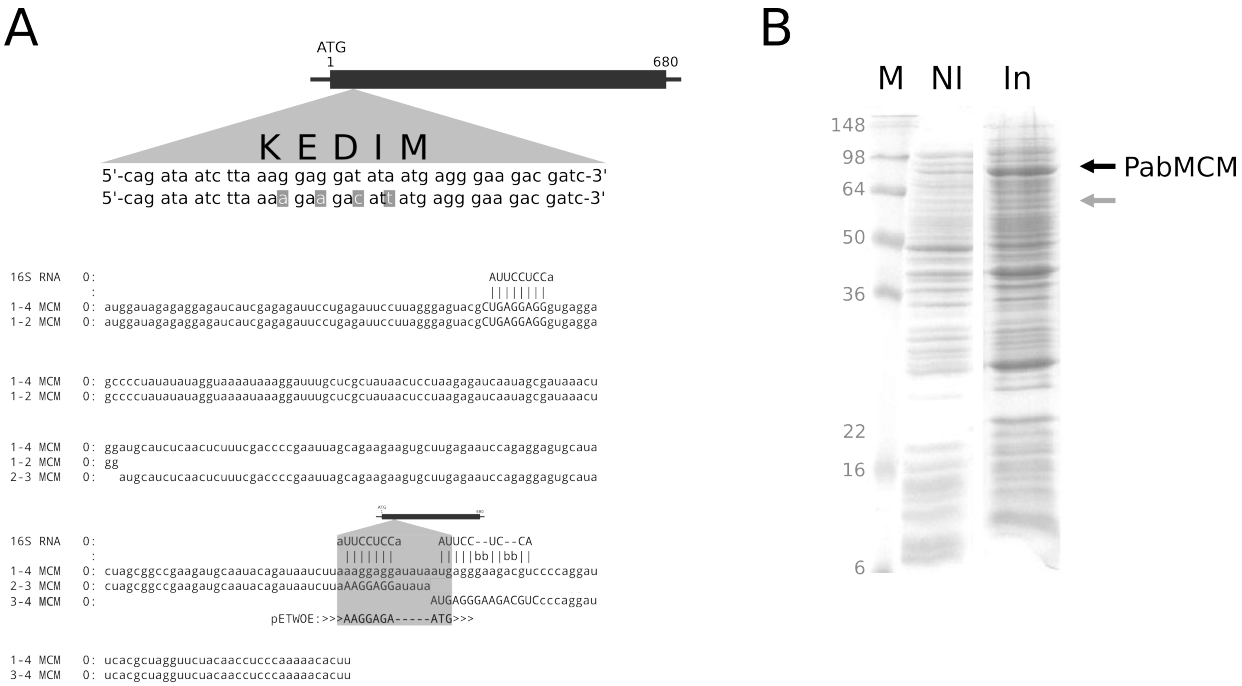



**Figure 4.12: Mutagenesis of the putative SD sequence of *Pab*MCM.** (A) Drawing of *Pab*MCM full-length with a closed-up of 5'–CAGATAATCTTAAAGGAGGATATAATGAGGGAAGACGATC–3' sequence mutated to 5'–CAGATAATCTTAAAAGAAGACATTATGAGGGAAGACGATC–3' sequence. Synonymous mutation of the amino–acidic sequence (KEDIM) was achieved via PCR-based method. Also, it is shown the results from Free2Bind. Sequences are aligned to the 306 nt sequence (1-4 MCM), which goes from the first methionine to the fourth one found in-frame in the sequence. 1-2 MCM is the nucleotide sequence from the first methionine to the second one. 2-3 MCM is the nucleotide sequence from the second methionine to the third one while 3-4 MCM is the nucleotide sequence from the third methionine to the fourth one. Small drawing highlight the putative alternative initiation of translation element found in the sequence. (B) SDS–PAGE of the mutated *Pab*MCM. NI, sample where protein expression was not induced. In sample where protein expression was induced with 1 mM IPTG. Clearly *Pab*MCM is expressed and the shorter *Pab*MCM fragment is certainly less.

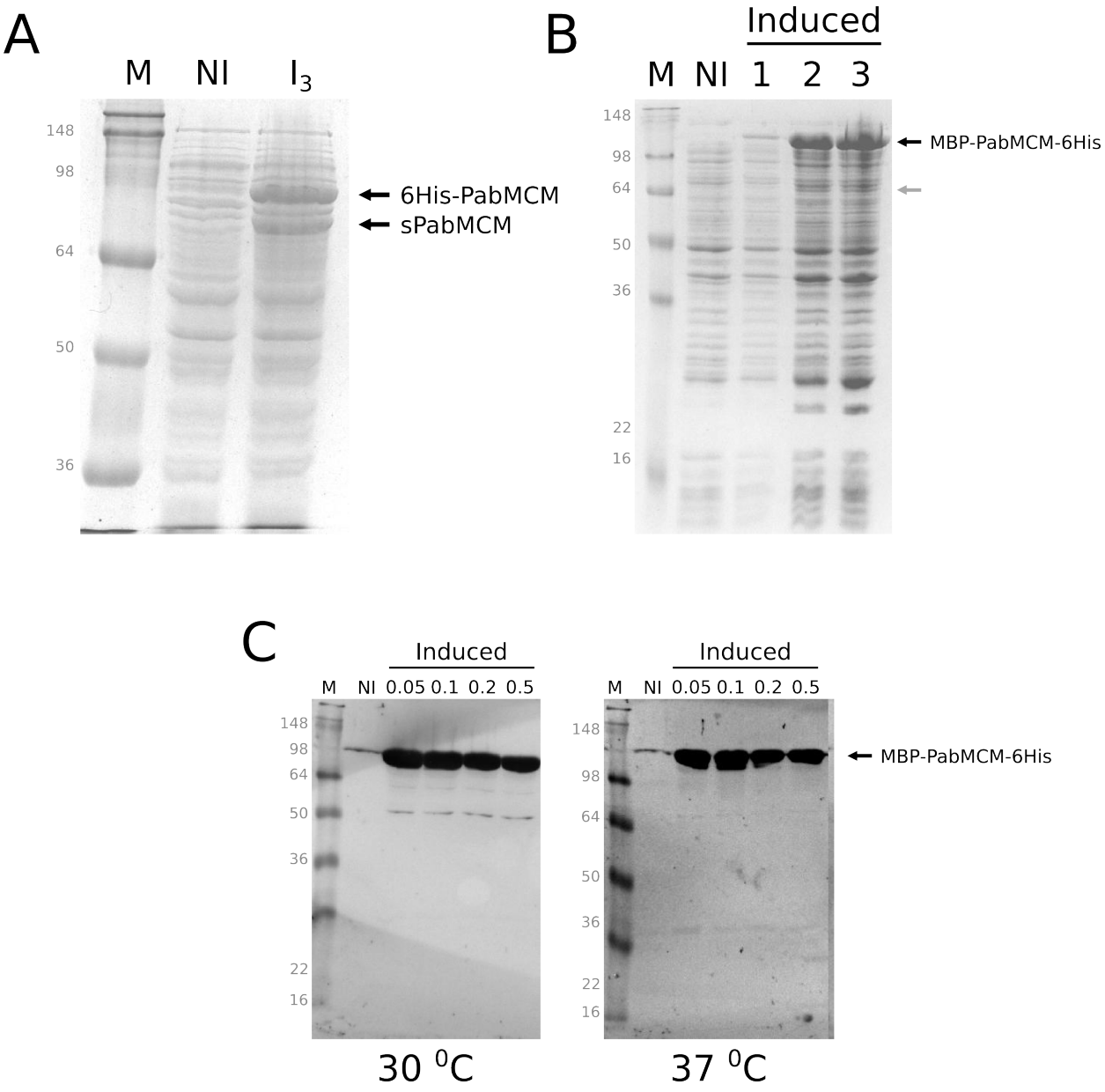



**Figure 4.13: Small scale protein expression test and western blotting analysis of the mutated *Pab*MCM.** (A) Shows the small protein expression test performed with the construct pTWOE-6HisMCMFL; NI, not induced sample while $I_3$ three hours post-induction cells with 1 mM IPTG. It is clearly seen the shorter *Pab*MCM fragments being co-expressed with the full-length one. (B) Shows the small protein expression test performed with the construct pETM40-MBP-MCMFL-6His; NI, not induced sample while 1, 2 and 3 post-induction samples after one, two and three hours of induction with 1 mM IPTG. It is clearly seen that the shorter *Pab*MCM fragments is not being co-expressed anymore with the full-length one. (C) Shows the western blotting of the small protein expression test performed with the construct pETM40-MBP-MCMFL-6His at 30°C and 37°C respectively; NI, not induced sample while 0.05, 0.1,0.2 and 0.5 are IPTG concentration used in this screening. Samples were induced over-night in TB. This figure shows clearly that the shorter *Pab*MCM fragments is not being co-expressed any more with the full-length one also it is show that neither proteolysis nor degradation is occurring after ∼16 hours post induction.

The third match ($\Delta$G = –7.9) also might be another alternative initiation of translation element for two reasons. Firstly, although the predicted binding affinity is not as high as the one for the second match there are still chances that the second match might be used as initiation of translation. Secondly, as reviewed in Romero and Garćia [1991] initiation of translation at AUC, AUA and AUU codons in *E. coli* might occur. Further analyses need to be implemented to address whether or not this codons are being used as alternative initiation of translation elements.

### 4.4.5 An optimised purification protocol for *Pab*MCM

All the experiments shown before have been used as lead to set a protocol for an heterologous protein expression. No improvement in the purity of the protein were noticed after heparin affinity and strong cation exchange chromatography when performed after nickel affinity chromatography (data not shown). Hence, the final optimised protocol was carried out in three steps, which included: heat–denaturation at 70°C, nickel affinity chromatography (Figure 4.14) and size–exclusion chromatography (Figure 4.15).

Eluted fractions from the nickel affinity purification step were pooled and checked for DNA contaminations measuring the $A_{260/280}$ ratio at a nanodrop spectrophotometer. Only fractions with a $A_{260/280}$ between 0.51 and 0.68 (99% purity) were concentrated by ultra–filtration with a 15R Vivaspin (15000 MWCO), to a final concentration of 15 mg/ml. Finally, sample was loaded onto a Superose6$^{TM}$10/300 GL size–exclusion column. The apparent molecular weight, estimated as in Andrew [1970], of *Pab*MCM was 575 kDa, indicating that it is more likely to be an octameric assembly.

## 4.5 DNA binding assay

Structural studies on archaeal MCM proteins have shown a positively charge central channel large enough to allocate dsDNA [Brewster et al., 2008; Fletcher et al., 2003]. In addition, biochemical studies reported that MCM proteins bind DNA through two types of structural motifs, which include a zing finger and a $\beta$–hairpin motif

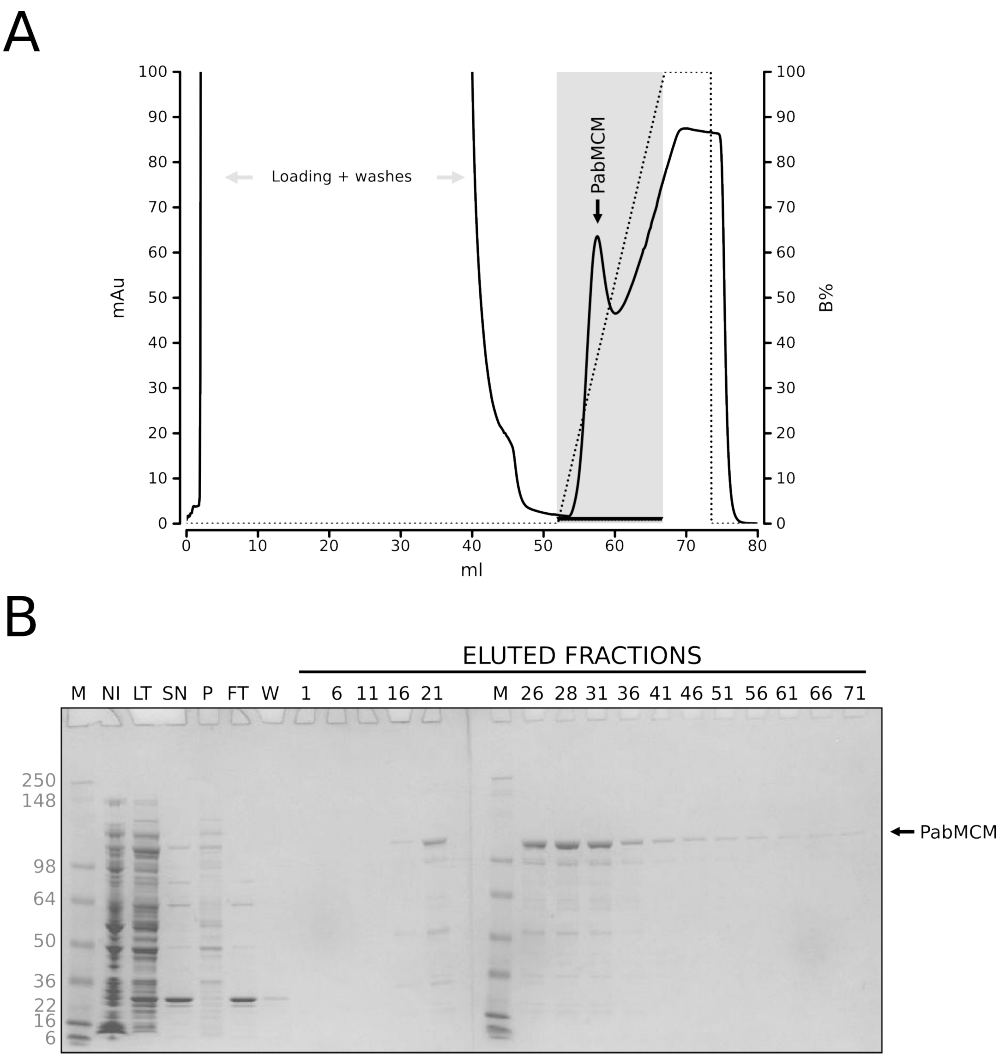



**Figure 4.14: Optimesed purification protocol for *Pab*MCM, Nickel affinity purification chromatography step.** (A) Nickel affinity purification chromatography trace. Black continuous line is the abosorbance measured at 260 nm (mAu) while dotted line is the gradient of buffer B (%B) used for eluting the 6HisMCMFLM recombinant protein. Sample in 20 mM HEPES pH 7.4, 500 mM NaCl, 20 mM imidazole was loaded onto an nickel 5 ml nickel column at 1 ml/min flow rate. Nickel bounded protein wash washed in 20 mM HEPES pH 7.4, 1.5 M NaCl and re–equilibrated in 20 mM HEPES pH 7.4, 150 mM NaCl, 20 mM imidazole. Finally, sample was eluted with 15 CV (75 ml) of 20 mM HEPES pH 7.4, 150 mM NaCl, 500 mM imidazole at 5 ml/min flow-rate. Shaded in light grey are the fractions collected during the chromatographic run. (B) NI, not induced Rosetta$^{TM}$(DE3) pLysS host cells harbouring pETWO-6HisMCMFLM construct were induced in TB at 30°C with 0.02 mM IPTG when $OD_{6}00$ reached 0.5 – 0.6; LT, total lysate of harvested cells broken by sonication at 50% amplitude in 5 cycles of 30" each with pause of 30" in between each cycle; SN, supernatant of heat treated sample at 70°Cfor 20' and loaded onto a 5 ml pre-packed nickel column; (P) Pellet after heat-denaturation; FT, flow-through after SN was loaded onto the nickel column; W, Nickel bounded protein washed with 1.5 M NaCl, which is now to induce disassembly of the ring in *S. solfataricus* MCM protein complex [Brewster et al., 2008]. This wash was included in order to remove any aspecific ionic interactions with the protein; ELUTED FRACTIONS, fractions collected after sample was eluted with a linear gradient of imidazole.

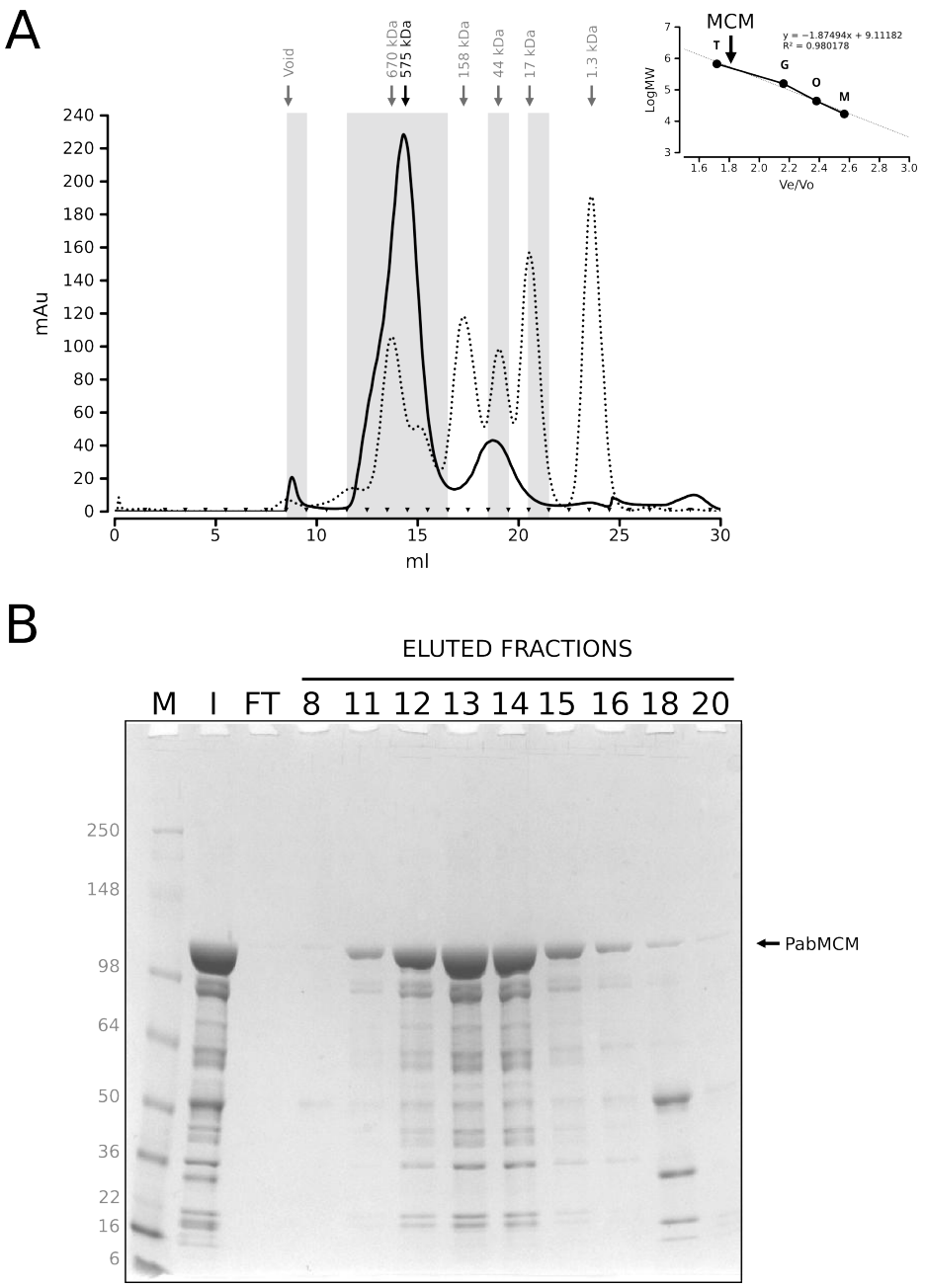



**Figure 4.15: Optimesed purification protocol for *Pab*MCM, Size-exclusion chromatography step.** Gel-filtration was performed in 20 mM HEPES pH 7.4, 150 mM NaCl at 0.3 flow-rate. (A) Black continuous line, size-exclusion chromatography trace of *Pab*MCM. Black dotted line, gel-filtration protein standard used for calibrating the column and estimate the apparent molecular weight of *Pab*MCM in solution. As it is shown *Pab*MCM elutes very close to thyroglobulin protein standard. Shaded in light grey are the fractions analysed on SDS–PAGE while arrows in light grey are the molecular weight of the protein standards used in this study. The black arrow indicates the estimated molecular weight of *Pab*MCM. Top left graph, shows the calibration curve used in this study (T = Thyroglobulin (670 kDa); G = $\gamma$-globulin (158 kDa); O = ovalbumin (44 kDa); M = myosin (17 kDa)). (B) SDS-PAGE analysis of the eluted fractions. I, sample input (diluted 10 times). FT, flow-through after ultra-filtration. ELUTED FRACTION, fraction collected during size-exclusion chromatography.

at the N–terminus and a $\beta$–hairpin located in the AAA$^+$ catalytic domain [Grainge et al., 2006; Kasiviswanathan et al., 2004; McGeoch et al., 2005]. Bioinformatics analysis shown a strong sequence homology, with well conserved structural motifs, with well–known archaeal MCM protein complexes (Figure 4.1). Supporting these findings, an agarose gel based shift assay was implemented in this part of this study to explore whether or not *Pab*MCM binds dsDNA. Indeed, the *Pab*MCM protein complex was found to bind dsDNA (Figure 4.16).

## 4.6 Concluding remarks

All archaeal genomes sequenced to date possess homologues of eukaryal MCM proteins [Barry and Bell, 2006]. Genome sequencing and bioinformatics analyses of hyperthermophilic archaeon *P. abyssi* revealed an ORF (PAB2373) with sequence homology to previously reported MCM proteins [Cohen et al., 2003].

Comparison of the primary amino acid sequence of ORF PAB2373 (herein *Pab*MCM) with three well–characterised archaeal MCM proteins revealed conserved residues in the N–terminal domain of *Pab*MCM. The C domain of the N–terminal portion of MCM proteins contains an important loop (Figure 4.1, residues 187–203; Figure 1.4), the N–terminal communication loop, which plays a key role in coordinating DNA binding and ATPase activity between subunits [Sakakibara et al., 2008; Barry et al., 2009]. Another well–conserved functional motif is the $\beta$–hairpin, which is involved in DNA binding [Fletcher et al., 2003]. Bioinformatic analyses of the C–terminal of *Pab*MCM show highly conserved residues (Figure 4.2). Functional motifs entailed in ATP binding and hydrolysis (Walker A and B, SI, SII and SRF) are all well conserved.

Bioinformatics analyses of the genomic sequence of *P. abyssi* reported the presence of two inteins in the *Pab*MCM coding sequence. The first intein insertion, breaks apart C–terminal of the Walker A motif, whilst the second intein insertion falls into the glutamate switch, close to the $\beta$–hairpin H2I (Figure 4.2). The glutamate switch senses the presence or absence of a ligand, which is DNA in this case [Zhang and Wigley, 2008].

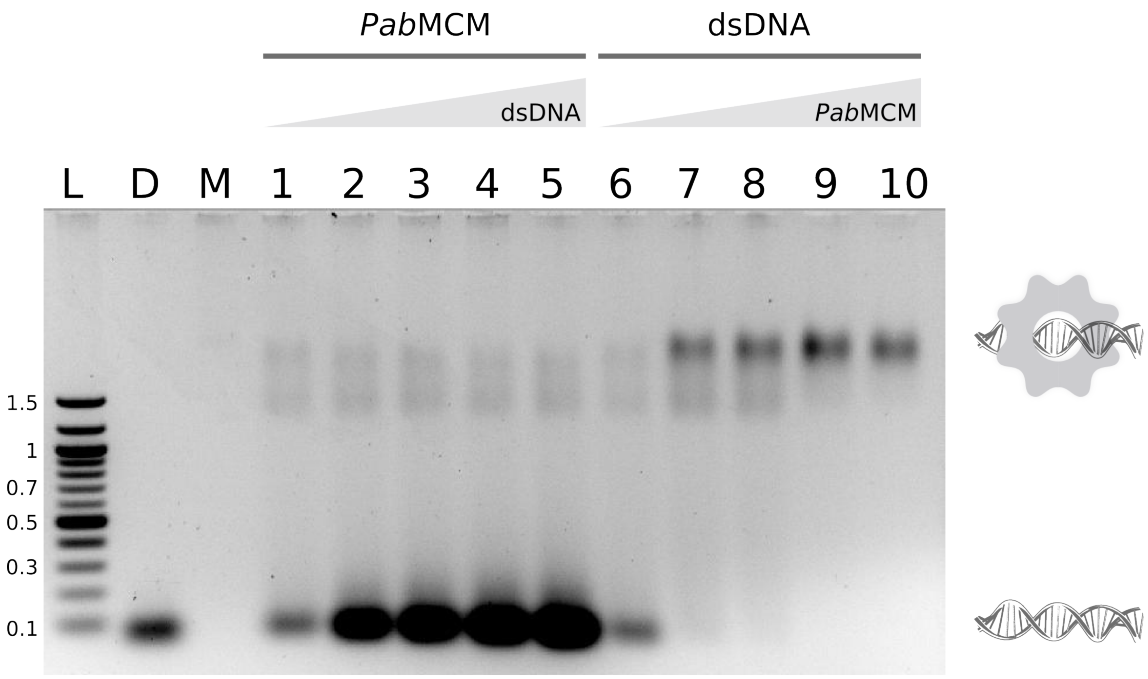



**Figure 4.16: DNA binding assay.** The gel shows the *Pab*MCM complex can bind dsDNA. A 59-mer dsDNA was used in this experiment. Size was chosen in order to avoid multi-MCMs loading on the same dsDNA and thus force a 1:1 stoichiometry. L, DNA ladder. D, dsDNA only used has control. M, *Pab*MCM only used as control to avoid artefacts due to ethidium bromide staining. Lines 1-5 all contain 24 pmol (calculated on the monomeric MCM) of *Pab*MCM with increasing amount of dsDNA (10, 20, 30, 40, 50 pmol). Lines 6-10 all contain 10 pmol of dsDNA with increasing amount of *Pab*MCM (24, 48, 72, 96, 120 pmol). The band-shift experiment was performed as shown in Fletcher et al. [2003].

Based on the observation of the presence of two inteins, in order to biochemically characterised the full–length *Pab*MCM the ORF2373 was reconstructed by a PCR–based approach (Figure 4.4). By this approach, the full–length coding sequence of *Pab*MCM was amplified and then cloned in an appropriate expression vector. The nucleotide sequence of *Pab*MCM was confirmed by sequencing (Figure 4.6). Preliminary over–expression tests, carried out in *E. coli* strains (Figure 4.7), revealed that *Pab*MCM was over–expressed, although a smaller polypeptide (60 kDa) was also co–expressed. Further heat denaturation tests (Figure 4.8) and chromatography binding assays showed that the *Pab*MCM protein was thermostable and could bind an cation exchange resin and an heparin affinity resin Figure 4.9). Several approaches were tested to separate the 60 kDa co–expressed protein and the full-length *Pab*MCM (Figure 4.10). Mass–spectrometry analyses revealed that, the 60 kDa co–purifying polypeptide was a shorter form of *Pab*MCM lacking the N–terminal domain Figure 4.11. In–depth analyses, of the 5'–end of the nucleotide sequence of *Pab*MCM, revealed the presence of a putative RBS element (Figure 4.12 A). Mutagenesis of this putative RBS element resulted in the absence of the co–expressed 60 kDa polypeptide (Figure 4.12 B). Finally, this mutated construct was used for setting up a purification protocol for the full-length *Pab*MCM (Figure 4.14, Figure 4.15).

DNA binding assays carried out with the mutated full-length *Pab*MCM showed that *Pab*MCM binds to dsDNA (Figure 4.16). In eukaryotes, the loading of MCM2–7 onto the DNA is mediated by ORC, Cdc6 and Cdt1 in a ATP–dependent process. In bacteria and virus, the replicative helicase is loaded onto the DNA by a loader protein or complex, which assembles the helicase onto the origin of replication [Sakakibara et al., 2009a]. In archaea, it is currently unknown whether the replicative helicase has loader proteins or not. However, due to the eukaryotic–like mode of replication it is possible that some proteins could act as loader for the replicative helicase. A candidate for this purpose is Cdc6 [Shin et al., 2008]. This assumption relies on the sequence similarity found between the Cdc6 in archaea and the eukaryotic initiator protein Cdc6 [De Felice et al., 2003].

# Chapter 5

# Results and Discussion: the structure of the archaeal *Pab*MCM protein complex

## 5.1 Introduction

The mini–chromosome maintenance (MCM) proteins are members of the $AAA^+$ superfamily of ATPases [Bell and Botchan, 2013]. $AAA^+$ ATPases use energy derived from ATP binding and hydrolysis to carry out specific functions in several biological processes (e.g. DNA replication and repair) [Ogura and Wilkinson, 2001b]. In the case of MCM proteins, the $AAA^+$ module is needed to catalyse DNA unwinding [Bell and Botchan, 2013]. MCMs also have an amino–terminal domain (NTD), which plays a role in higher order structure assembly [Sakakibara et al., 2009b]. Following the $AAA^+$ domain, there is a winged helix (WH) motif, which may have a regulative role (Figure 1.4) [Barry et al., 2007; Jenkinson and Chong, 2006; Fernández-Cid et al., 2013].

Much of what we know regarding MCM helicases has been gathered from structural and functional studies of the simple archaeal model [Sakakibara et al., 2009b]. In addition, the crystal structures of distantly related helicases (SV40, LTAg, and E1 helicase of bovine papilloma virus) have been source of important structural frameworks for understanding the mode of action of hexameric helicases [Bell and

Botchan, 2013].

The first structural observation of a higher order assembly of MCM proteins was from electron microscopy studies of *M. thermautotrophicus* MCM [Bell and Botchan, 2013]. These studies showed a double hexameric structure with the two hexamers joined in a head–to–head manner [Chong et al., 2000], similar to the MCM2-7 assembly observed later [Remus et al., 2009]. Subsequently, it was shown that more oligomeric states could be adopted by MCM proteins, including single hexamer, single heptamer, double heptamer and filaments [Yu et al., 2002; Pape et al., 2003; Chen et al., 2005; Costa et al., 2006b; Gómez-Llorente et al., 2005; Bae et al., 2009; Slaymaker et al., 2013].

## 5.2 Electron microscopy and single particle analysis

To glean insights into the three–dimensional architecture of the *Pab*MCM complex, electron microscopy coupled to single particle analysis (SPA) experiments were performed. Negative stain electron microscopy images of the full-length *Pab*MCM were used for SPA. A typical micrograph is shown in Figure 5.1 A. ∼120,000 molecular images (320 x 320 pixels, 1.51 Å/pixel) were pre–processed as explained in section 3.9. The final size of the molecular images used for during data processing was 80 x 80 pixels at 6.04 Å/pixel.

A first round of MSA classification, which consists of a first MSA and a subsequent classification, was performed in order to calculate reference–free class–averages and build a first catalogue of views present in the dataset.

The resulting eigenimages, calculated by MSA, are shown in Figure 5.2. The eigenimages are normally presented in the order of their significance [White et al., 2004]. In this respect, eigenimage 2 reflects the characteristic of the main shape,which is a ring, in the dataset. Additionally, the strong black ring encircling the white ring is likely to be related to the size variation between particles into the dataset [White et al., 2004; Morris et al., 2011]. Eigenimages 3 and 4 have two–fold symmetry, which is likely to be related to a different oligomeric states (monomer/dimer) [Morris et al., 2011]. Eigenimages 5 and 6 reflect size variation

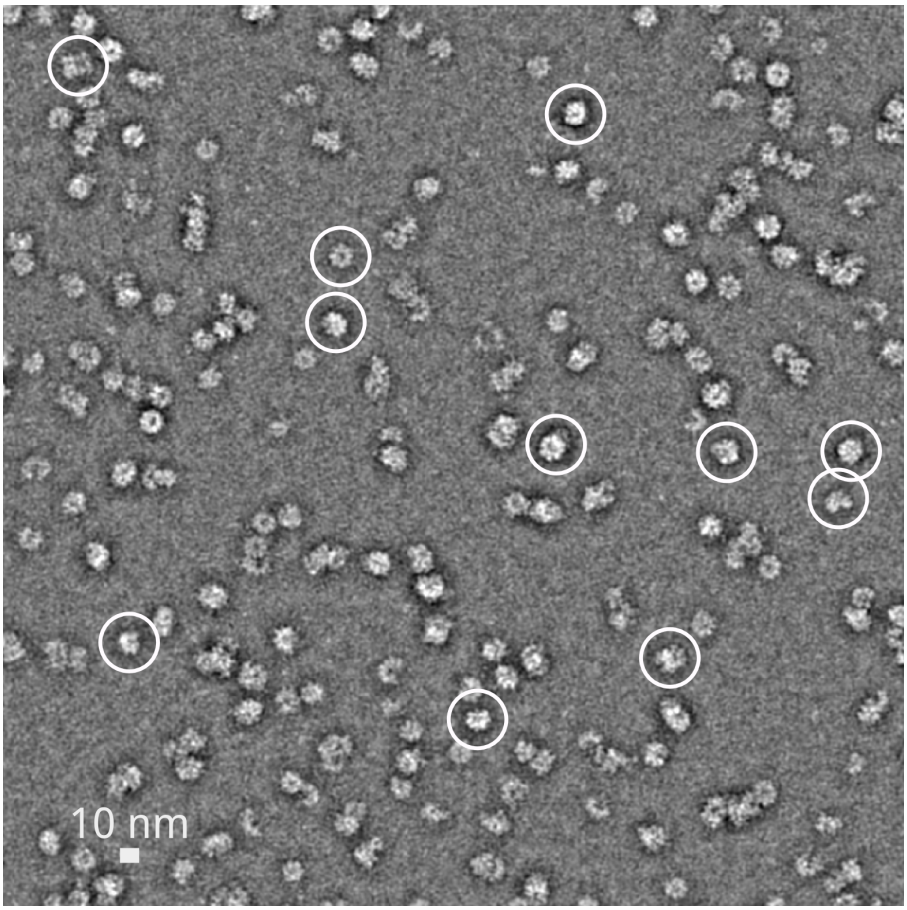



**Figure 5.1: Characteristic negatively stained electron micrograph of *Pab*MCM.** Micrography recorded at 50000x nominal magnification in low–dose mode (20–25 $e^-/Å^2$.). White circles are used to show single particles.

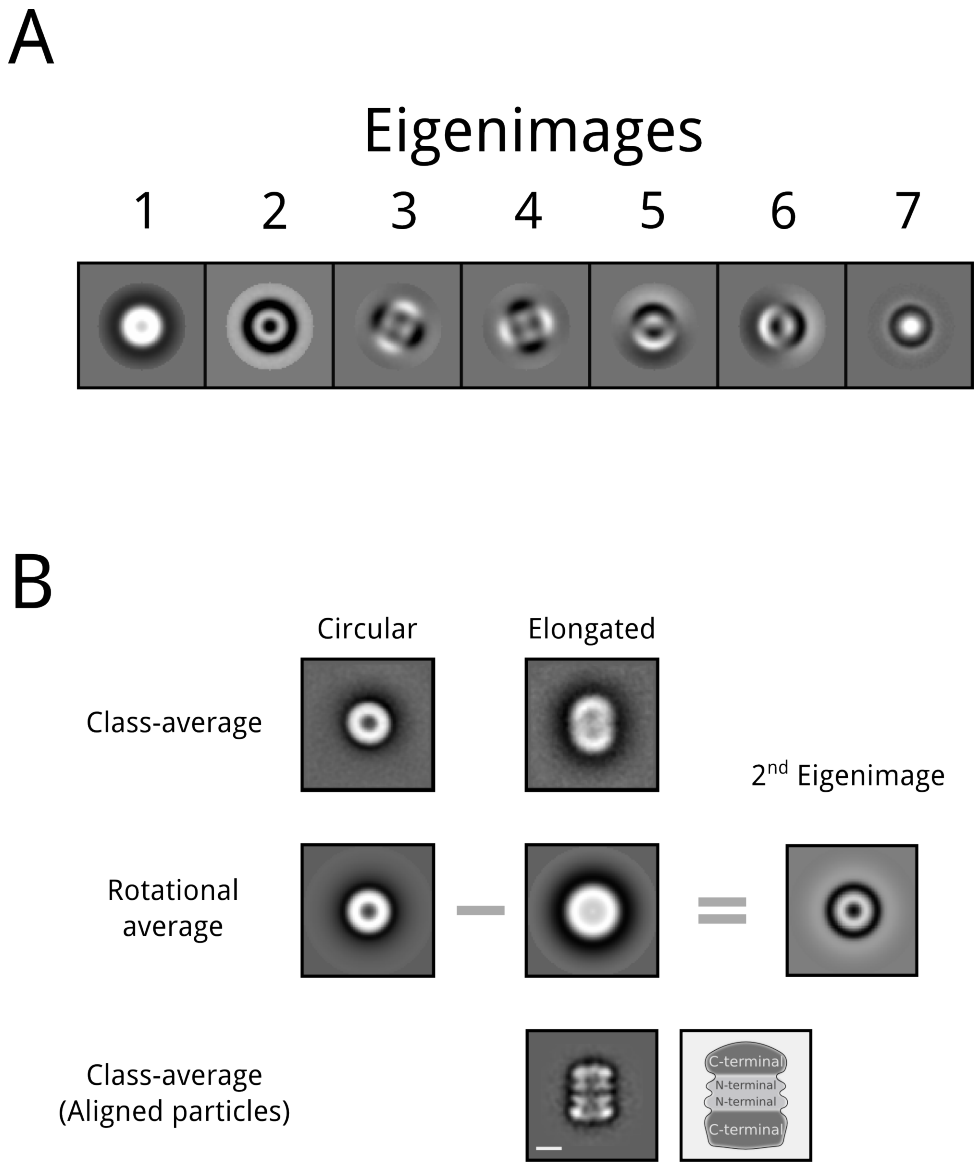



**Figure 5.2: Initial MSA classification of *Pab*MCM molecular images (part I).** (A) First eigenimage used for classification of the molecular images. (B) Two characteristics class-average, which show size variation. Class–averages were compared by subtraction (Circular minus elongated) to verify that the eigenimage 2 was representative of the difference in size between circular and elongated particles. Further MSA classification followed by multi–reference alignment (MRA) of the molecular images belonging to the elongated class–average revealed a characteristic two–fold symmetry and four tiers structure. These features are similar to those ones previously observed for *Mth*MCM. Scale bar 100 Å.

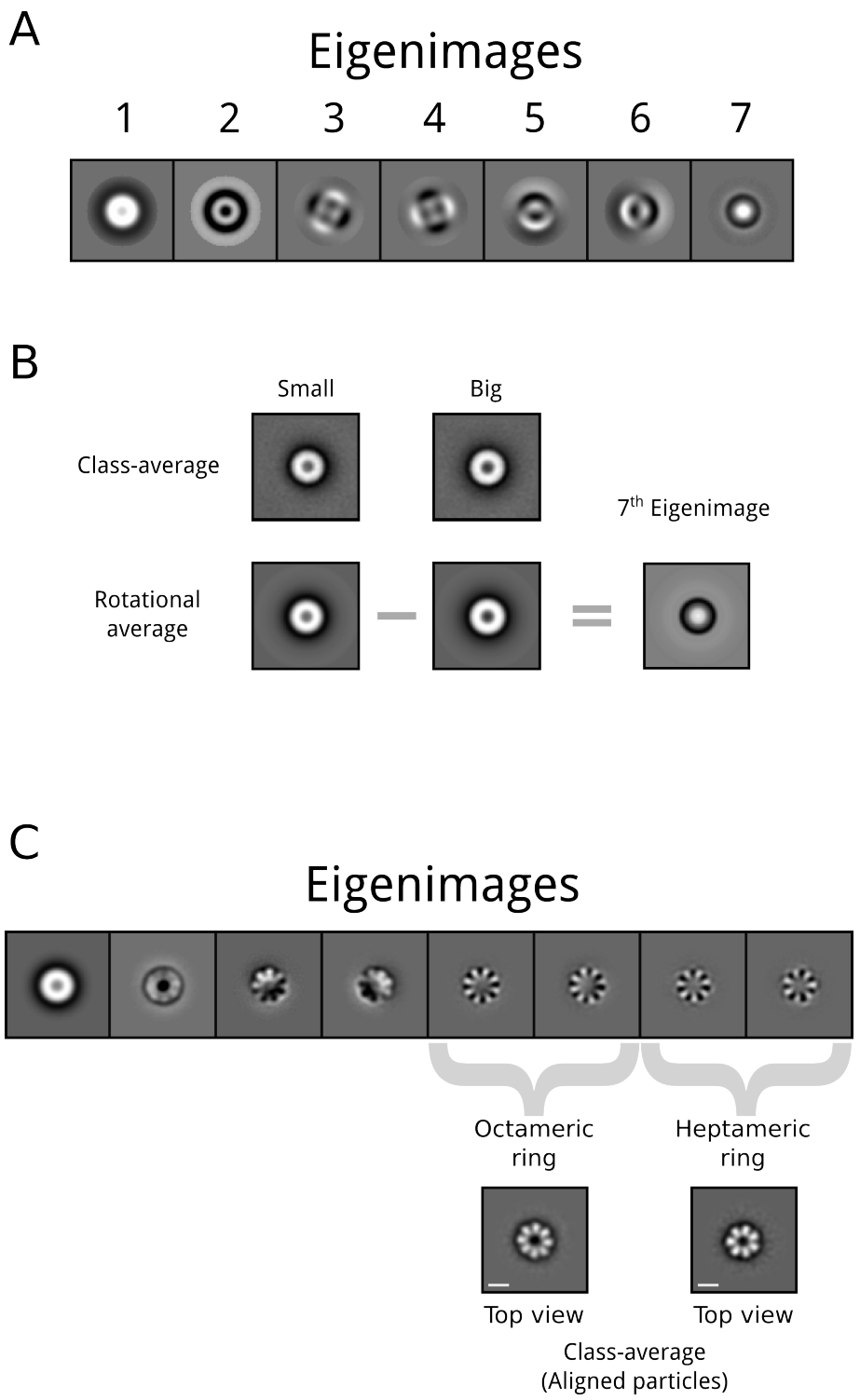



**Figure 5.3: Initial MSA classification of *Pab*MCM molecular images (part II).** (A) First eigenimage used for classification of the molecular images. (B) Two characteristics class-average, which show size variation. Class–averages were compared by subtraction (Small minus big) to verify that the eigenimage 7 was representative of the difference in size between small and big particles. (C) Further MSA classification followed by multi–reference alignment (MRA) of the molecular images belonging to the both class–averages revealed a two ring–shaped class–averaged with a characteristic 8–fold and 7–fold symmetry for *Mth*MCM. Scale bar 100 Å.

for the tilted views. Eigenimage 7 shows a black and a white rings, probably still related to the different particles size within the dataset [White et al., 2004]. To check whether or not this interpretation was correct, molecular images were classified using the first seven eigenimages. 100 class–averages were calculated (data not shown). Analysis of these class-averages showed three types of class-averages: an elongated (Figure 5.2) and two ring–shaped (Figure 5.3). The validity of the interpretation of the $2^{nd}$ and $7^{th}$ eigenimage was checked in a similar way as shown in [White et al., 2004]. Rotational averages of the circular and elongated class–averages were calculated and then images were compared by subtraction. As shown in Figure 5.2 and Figure 5.3 both eigenimages are representative of the size variation in the dataset. Based on this observation the dataset was partitioned into two sub–datasets: the elongated and the ring–shaped. Further rounds of MSA classification and alignment, with a subset of ∼5000 molecular images classified as elongated, revealed their features. As shown in Figure 5.2 the elongated class–average shows a two–fold symmetry and a four–tier organisation, which is the typical feature of a double ring assembly [Costa et al., 2006b; Gómez-Llorente et al., 2005]. This is consistent with previous electron microscopy studies of *Mth*MCM [Costa et al., 2006b; Gómez-Llorente et al., 2005]. Modelling the EM structure of *Mth*MCM double hexamer indicated that the top and bottom tiers correspond to the C–termini, while the two middle tiers correspond to the N–termini (Figure 5.2).

Analogous analyses were carried out for the 45,000 image–subset of ring–shaped molecular images. As shown in Figure 5.3, further rounds of MSA classification leaded to classify two types of top-end views of *Pab*MCM: a small ring with a characteristic 7–fold symmetry and a bigger ring with a characteristic 8–fold symmetry. The difference in size estimated between the two rings was 8%. MSA classification also allowed to sort two types of side views, small and large side views (data not shown). More difficult was the sorting of the tilted views since the difference in size was not always obvious. However, those tilted views, which clearly belonged to the larger *Pab*MCM assembly were sorted manually. The goodness of these tilted view was validated in further steps during refinement of the final 3D model of the octameric *Pab*MCM. Finally, from this sub dataset, ∼10,000 molecular images be-

longing to the 'large' *Pab*MCM assembly were taken forward for calculating the initial 3D model of the full-length *Pab*MCM (Figure 5.4).

## 5.3 3D EM reconstruction of the archaeal *Pab*MCM

~10,000 molecular images, sorted by classification, were used to calculate and refine the final 3D volume in EMAN2 [Tang et al., 2007] (Figure 5.4, Figure 5.5). The initial model was calculated with `e2initial_moldel.py`, which calculates a random blob model, from pure noise images, to seed a single particle reconstruction and refinement [Tang et al., 2007]. For this purpose, nine class–averages were used for the 3D reconstruction of the initial model (Figure 5.4). In particular, ten initial models were calculated. For each 3D reconstruction, the `e2initial_moldel.py` program calculates a set of re-projections, which can be used to estimate 'how good' is the 3D reconstruction. Indeed, each 3D reconstruction was checked by looking at the 3D model and the correspondence between the class–averages and the re–projections. The best model, in which case was the second reconstruction was chosen since the best matching between class–averages and re–projections (Figure 5.4). This initial model was then refined, by projection matching with `e2refine.py` [Tang et al., 2007]. Re–projections and class–averages were checked at each iteration. Particular care was taken for the tilted views. As mentioned before, It was not straightforward to distinguish between small and large tilted views during MSA classification. This was taken into account during each iteration during refinement and 'bad' ones were discarded. The final refined model is presented in Figure 5.5 A. Projections of the maps matched with 2D class-averages assigned the same Euler angles highlighting the validity of the map (Figure 5.5 B). The overall shape of the refined model *Pab*MCM exhibits similarity with 3D–EM models previously reported archaeal MCMs [Pape et al., 2003; Bae et al., 2009]. The *Pab*MCM complex exhibits an octameric assembly and has overall dimensions of 170 Å x 170 Å x 110 Å. The resolution of the map is 22 Å, calculated at 0.143 Fourier shell correlation (FSC) (Figure 5.5 C).

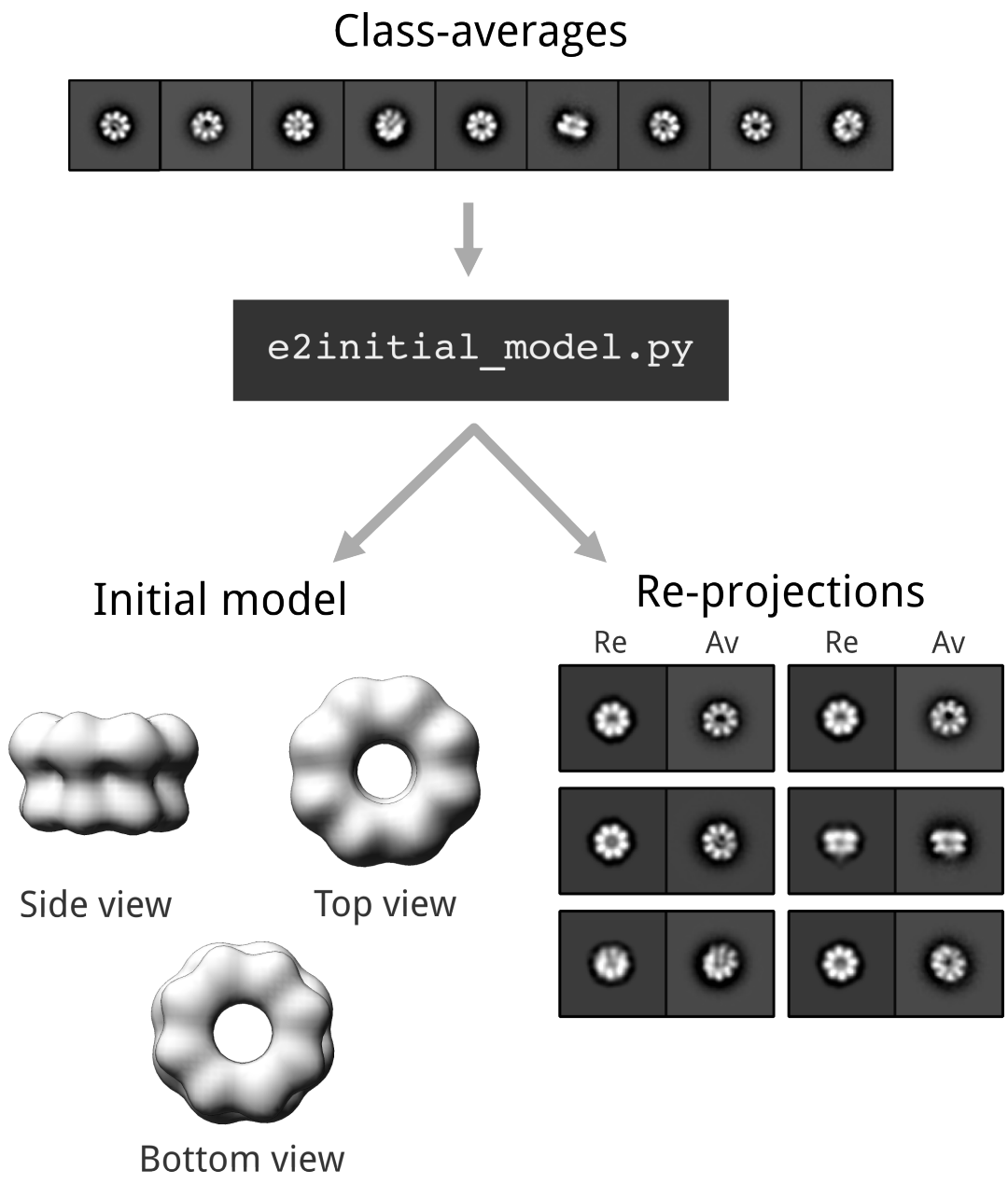



**Figure 5.4: Flow–through of the 3D reconstruction of the initial model of the full–length *Pab*MCM** From the top, 9 class–averages were selected for the 3D reconstruction of an initial model of the full–length *Pab*MCM. The initial model was calculated by using the `e2initial_model.py` software from the EMAN2 image processing suite [Tang et al., 2007]. The best initial model, showed in figure, was chosen based by evaluating the correspondence between class–averages (Av) and the re–projections (Re). The initial 3D model of the full–length *Pab*MCM single octamer was rendered in Chimera [Pettersen et al., 2004].

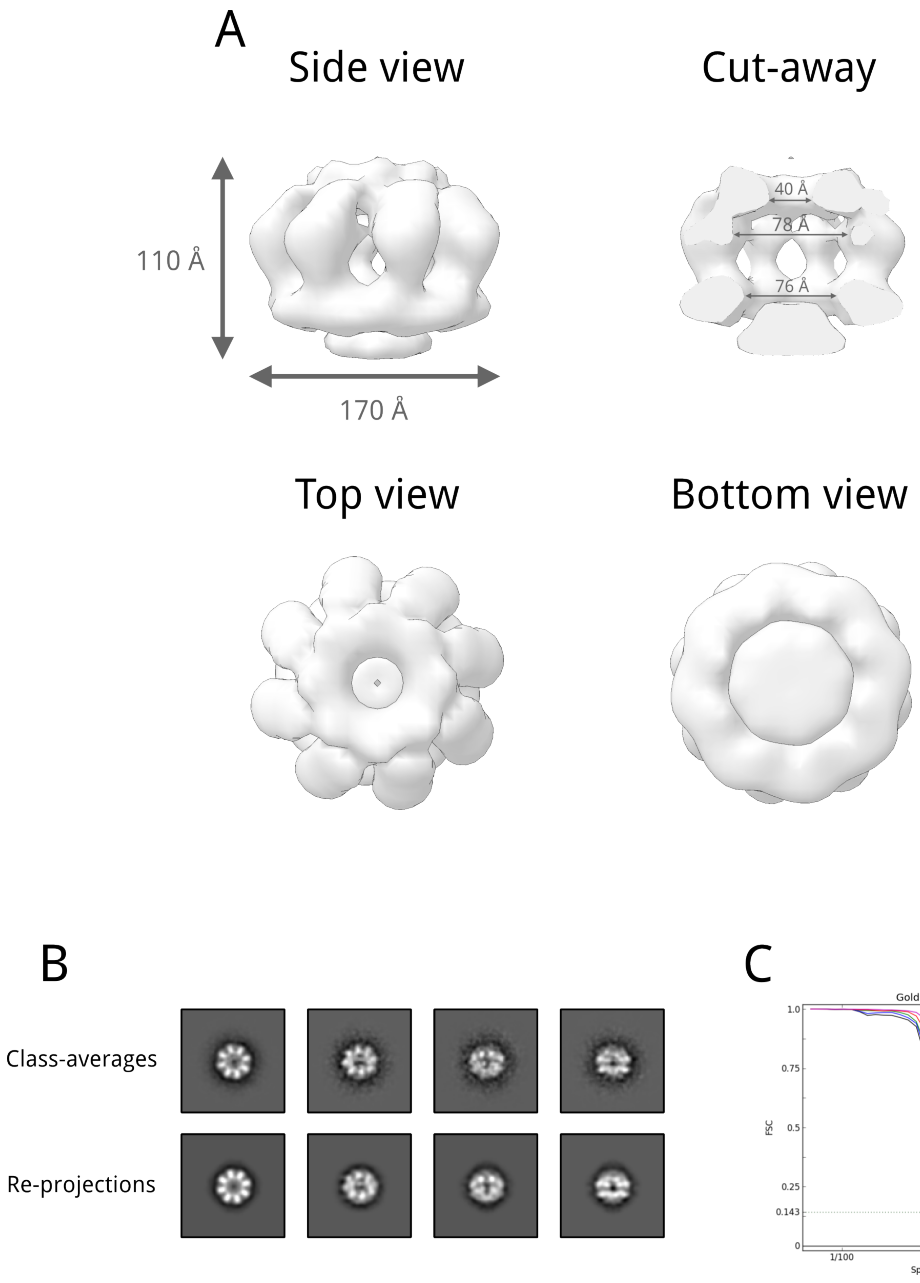



**Figure 5.5: 3D refinement of the full–length *Pab*MCM.** (A) Refined 3D model of the full–length *Pab*MCM single octamer. Volumes were rendered with Chimera [Pettersen et al., 2004]. (B) Class–averages and re–projections for the refined 3D reconstruction. (C) Fourier shell correlation of the refined *Pab*MCM model.

## 5.4 Model fitting of *Pab*MCM 3D–EM structure

To interpret the 3D–EM reconstruction of the full–length *Pab*MCM, docking of the atomic coordinates of the crystal structure C–terminal truncated *Sso*MCM [Brewster et al., 2008], the crystal structure of the full-length *Mka*MCM [Bae et al., 2009] and the NMR structure of the C–terminal domain of *Sso*MCM [Wiedemann et al., 2014] was performed . The fitting of the 3D structure of the full–length *Pab*MCM complex (Figure 5.6) was performed manually using Chimera [Pettersen et al., 2004] and then optimised via Situs [Wriggers et al., 1999]. 3D volume and fitted model were rendered in PyMol [DeLano, 2002].

The crystallographic structure of C–terminal truncated *Sso*MCM fitted well into the electron density corresponding to the bottom and top tier, although a C–terminal extra density was observed (Figure 5.6 A). The orientation resulting from docking of *Sso*MCM into the map shows PS1 and HP2 $\beta$–hairpins pointing into the central channel (Figure 5.6 B) while Ext $\beta$–hairpin is pointing toward the side channel between the subunits composing the ring (Figure 5.6 C). Ext $\beta$–hairpin locates on the exterior side of the side channel (close–up Figure 5.6 E).

The crystallographic structure of full-length *Mka*MCM fitted better into the electron density corresponding to the bottom and top tier (Figure 5.6 F). Docking resulted in PS1 and HP2 $\beta$–hairpins of *Mka*MCM pointing into the central channel of the 3D–EM map of the full-length *Pab*MCM (Figure 5.6 G, H and I). The overall shape of the 3D–EM reconstruction of full-length *Pab*MCM is very similar to the 18 Å cryo–EM map of *Mka*MCM complex [Bae et al., 2009].

## 5.5 Concluding remarks

The structural observation of archaeal MCM proteins by electron microscopy showed a double hexameric structure with the two hexameric rings joined in a head–to–head manner [Chong et al., 2000; Remus et al., 2009]. It was also shown that more oligomeric states could be adopted by archaeal MCM proteins, including single hexamer, single heptamer, double heptamer and filaments [Yu et al., 2002; Pape et al., 2003; Chen et al., 2005; Costa et al., 2006b; Gómez-Llorente et al., 2005; Bae

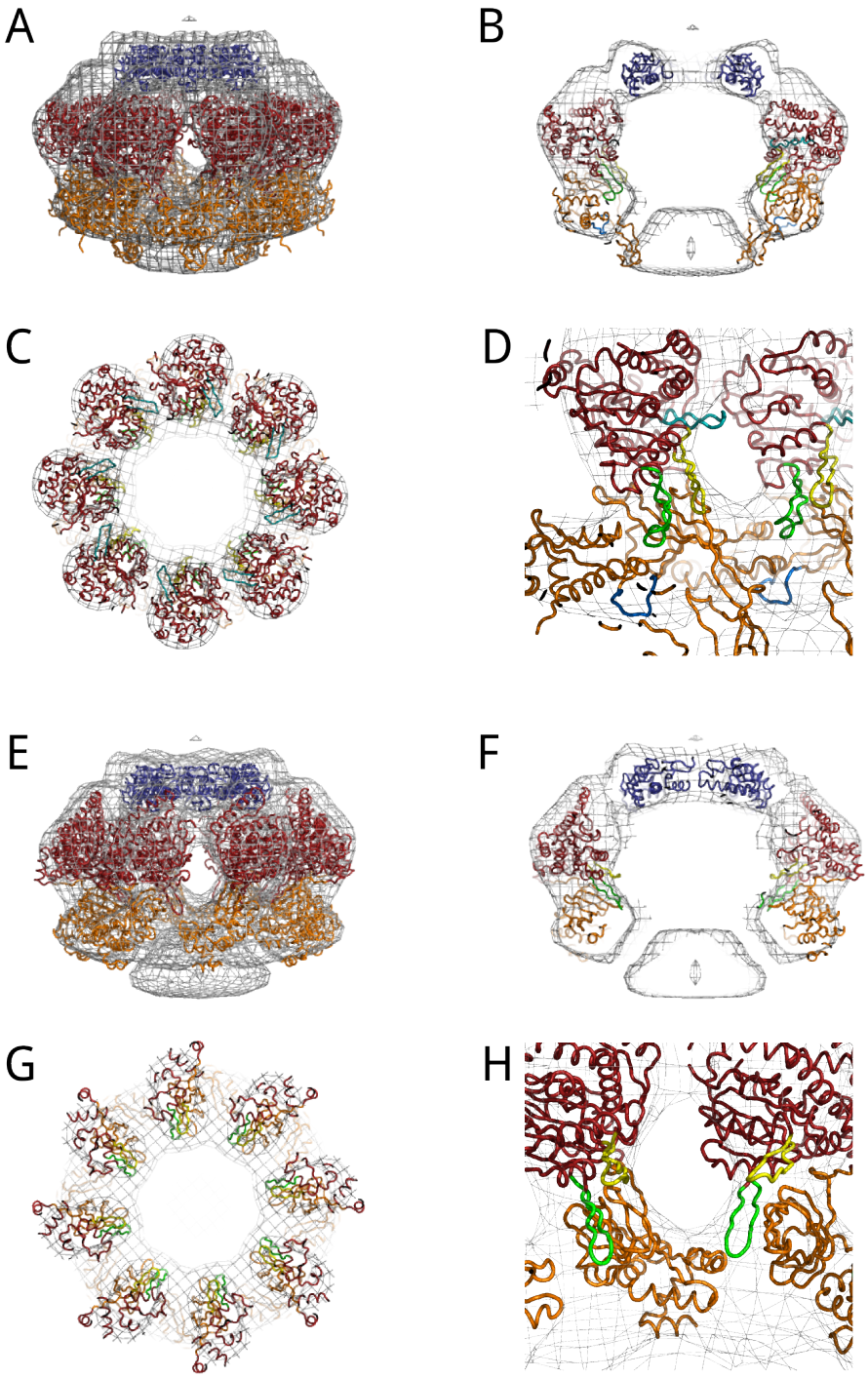



**Figure 5.6: Model fitting of the full–length *Pab*MCM 3D EM structure.** (A, B, C, D) Fitting of the atomic coordinates of the C–terminally truncated *Sso*MCM [Brewster et al., 2008] (PDB 3F9V) and C–terminal domain of *Sso*MCM (PDB 2M45)[Wiedemann et al., 2014]. (E, F, G, H) Fitting of the atomic coordinates of full–length *Mka*MCM [Bae et al., 2009] (PDB 3F8T). In blue is shown the fitting of the NMR structure of *Sso*MCM. In red, the AAA$^+$ module while in orange the N–terminal domain of both *Sso*MCM and *Mka*MCM. In light blue, EXT $\beta$–hairpin; in yellow, H2I $\beta$–hairpin; in green, PS1 $\beta$–hairpin; in dark blue, NT $\beta$–hairpin (only in panel D).

et al., 2009; Slaymaker et al., 2013].

In this study a homo–octameric assembly has been reported. SPA revealed that *Pab*MCM in solution is a mixture of at least three molecular species: single homo–heptamer, single homo–octamer and double rings. A subset of ∼10,000 particles, showing characteristic 8–fold symmetry and two–tiers features, were classified and sorted by MSA. This subset resulted to be homogeneous enough for a subsequent 3D reconstruction. The final 3D–EM map 22 Å showed similarity with previously reported 3D–EM reconstruction of similar archeal MCM proteins.

# Chapter 6

# Results and Discussion: small–angle scattering studies of the archaeal *Pab*MCM

## 6.1 Introduction

Small–angle scattering (SAS) is a powerful diffraction technique used for investigating the structure of matter [Feigin et al., 1987]. SAS is widely used in several branches of science, including structural and molecular biology [Feigin et al., 1987]. SAS is very powerful for studying large–scale structures up to 1 $\mu$m. By SAS one can gather information about the size and shape of structures in a sample [Feigin et al., 1987]. In SAS experiments, samples are exposed to a collimated beam of radiation (X–rays or neutrons). The radiation source is deflected by the interaction of it with the atomic structure of matter. The deflection or scattering pattern is recorded on a detector and then analysed. The deflection recorder is typically between 0.3° and 5° [Feigin et al., 1987]. Depending on the nature of the radiation used in SAS experiments, it can refer to: small–angle X–ray scattering (SAXS) or small–angle neutron scattering (SANS) [Feigin et al., 1987]. In SAXS experiments, X–rays with wavelength in the range of 0.5–2 Å are used to study the three dimensional structure of a sample; whereas in SANS experiments, thermal neutron in the range of 1–10 Å are used instead. In both cases, the scattering intensity I(s) is recorded as

a function of momentum transfer q (q=$4\pi\sin\theta/\lambda$, where $2\theta$ is the angle between the incident and scattered radiation and $\lambda$ the wavelength of the radiation) [Feigin et al., 1987]. The observed scattered intensity is the Fourier transform of the object shape in real–space and it can be expressed as follow:

$$\textbf{Scattered Intensity} \quad I(q) = N_p V_p^2 (\rho_p - \rho_s)^2 F(q) S(q) + B$$

where: $N_p$ = number of particle

$V_p$ = volume of particle

$\rho_p$ = scattering length density of particle

(electron density for X–rays and nuclear/spin density for neutrons)

$\rho_s$ = scattering length density of solvent

$F(q)$ = form factor

$S(q)$ = structure factor

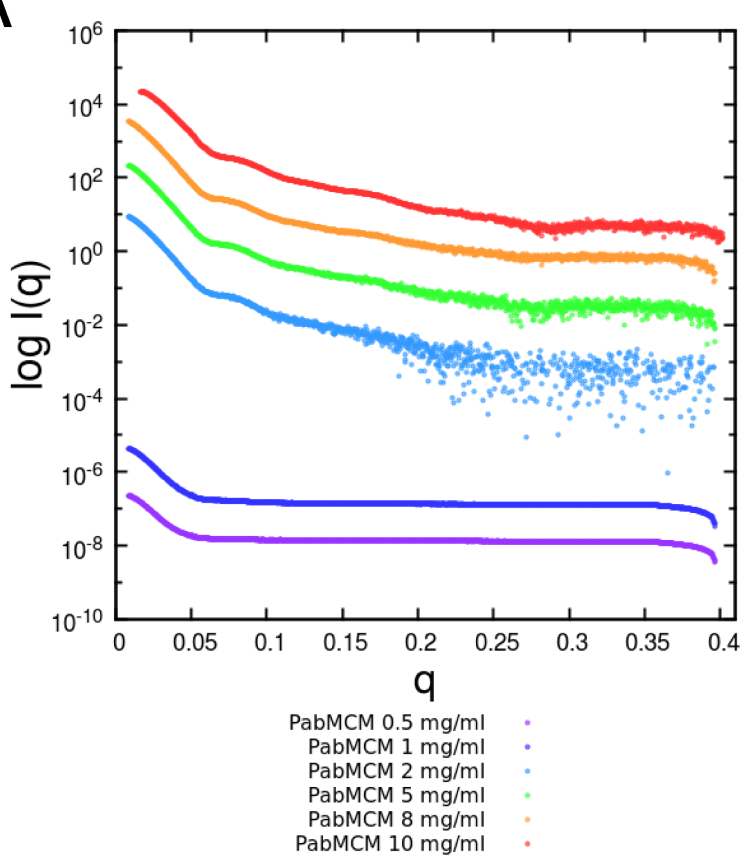
$B$ = background

Two considerations can be drawn from this equation [Edler, 2013]. Firstly, I(s) is directly proportional to the concentration ($I(q) \propto N_p V_p$), hence more sample more signal; however if too concentrated, inter–particles effect may occur [Jacques and Trewhella, 2010]. Secondly I(s) is directly proportional to difference in contrast between the particle and the solvent ($I(q) \propto (\rho_p - \rho_s)^2$), hence bigger particle scatter more than smaller one [Edler, 2013]. Importantly, the contrast can be varied in SANS using $H_2O/D_2O$ mixtures or selective deuteration in so called contrast match experiments [Jacques and Trewhella, 2010]. Hence, although both techniques are complementary, SANS over SAXS may be more informative since it is possible to study DNA–protein and protein complexes [Jacques and Trewhella, 2010].

## 6.2 Data analysis of *Pab*MCM SAXS curves

The three dimensional structure of *Pab*MCM full-length was investigated by SAXS. *Pab*MCM was prepared as shown in section 4.4.5. SAXS measurements were conducted at the beamline B21 at the Diamond Light Source (Harwell Science and Innovation Campus, Diamond, Oxford). Experimental buffer subtracted curves are shown in Figure 6.1. Data were collected at concentrations ranging from 0.5 mg/ml

A




B

| Pr. Conc. (mg/ml) | $R_g$(Å) | I(0) |
|---|---|---|
| 0.5 | 80 ± 0.1 | 0.27 ± 0.01 |
| 1 | 82 ± 0.1 | 0.5 ± 0.01 |
| 2 | 89 ± 0.1 | 1 ± 0.01 |
| 5 | 90 ± 0.1 | 2.6 ± 0.01 |
| 8 | 89 ± 0.1 | 4.1 ± 0.01 |
| 10 | 55 ± 0.1 | 3 ± 0.01 |

**Figure 6.1: Experimental buffer subtracted SAXS curves for *Pab*MCM.** (A) Buffer subtracted scattering curves of *Pab*MCM (Intensities (I(q)) are plotted in logarithmic scale while momentum transfer (q) is in $Å^{-1}$). Curves are offset to display features. Experimental solution scattering curves are shown separately to better display features. (B) $R_g$ an I(0) calculate by Guinier approximation.

to 10 mg/ml. Low protein concentrations (0.5–2 mg/ml) were used to avoid any possible aggregation and inter–particles effect, whereas to improve signal–to–noise at high angles, higher protein concentrations (5–10 mg/ml) were used instead. When curves for low protein concentration (0.5–1 mg/ml) were buffer subtracted, most of the signal for high angles was lost, although signal was still detected in the Guinier region (Figure 6.1 A). On the other hand, at concentration higher the 8 mg/ml inter–particles effect was quite severe. In fact, as shown in Figure 6.1, at 10 mg/ml the shape of the scattering curve in the Guinier region is more shallow. This was also confirmed by Guinier approximation since the $R_g$ estimated for this curve was smaller compared to the others [Jacques and Trewhella, 2010].

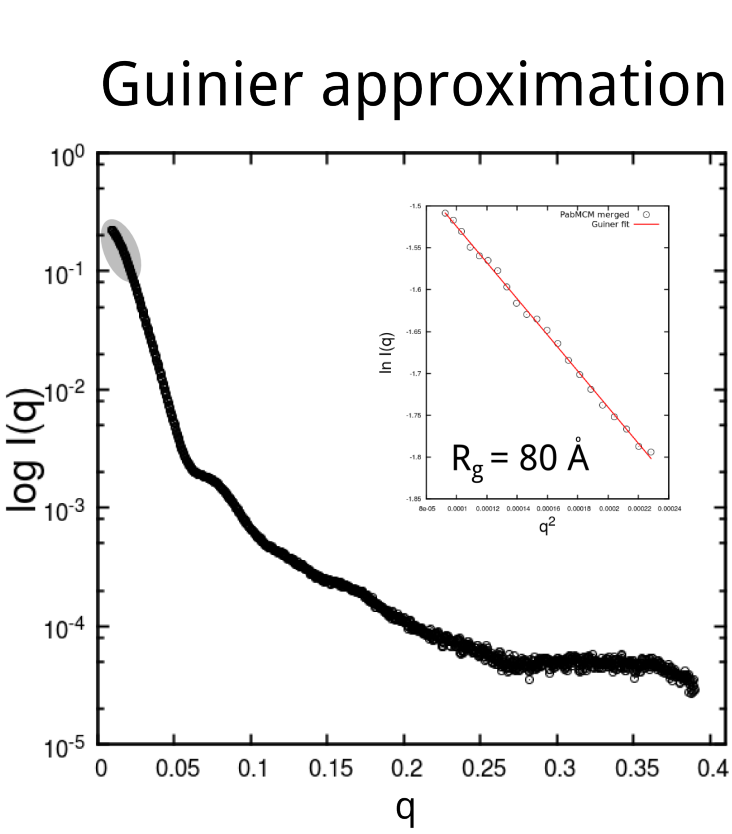
All the curves collected were scaled and merged into a final curve to make a concentration–independent scattering curve. Scaling and merging was achieved using PRIMUS [Konarev et al., 2003]. During merging, extreme care was taken to find overlapping points between the curves. The final merged curve (Figure 6.2) was used to estimate invariants ($R_g$, I(0) and $D_{max}$) by Guiner analysis and P(r) distribution function.

**Guinier approximation** A scattering curve contains structural information related to the shape and the size of the macromolecule in solution [Feigin et al., 1987].

Guinier approximation is based on the observation, by Andre Guinier, that at low concentration and very small scattering angles I(q) of a SAS scattering curve can be expressed as follow:

$$\textbf{Guinier Approximation} \quad I(q) = I(0)e^{\frac{-q^2R_g^2}{3}}$$

Structural parameters, $R_g$ and I(0), related to the size and shape of the scattering particle can be calculated by Guiner approximation [Guinier et al., 1955]. The $R_g$ (radius of gyration) is defined as the root-mean-squared distance of all elemental scattering volumes from their centre of mass weighted by their scattering densities [Jacques and Trewhella, 2010]. $R_g$ gives information about the mass distribution within a particle. Hence objects with the same volume but different shapes have different $R_g$ [Jacques and Trewhella, 2010]. I(0) (forward scattering intensity) is



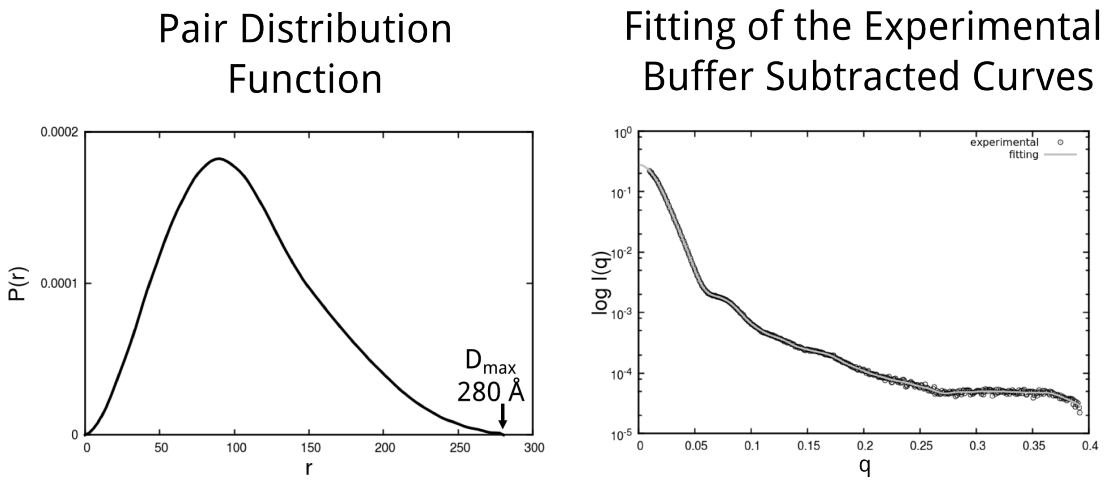




**Figure 6.2: SAXS data analysis of the *Pab*MCM merged curve.** Top panel, Guinier approximation of the merged scattering curve of *Pab*MCM (Intensities (I(q)) are plotted in logarithmic scale while momentum transfer (q) is in $Å^{-1}$). Grey oval indicates the Guinier region while inner plot the linear fitting of 21 point within the Guinier region. Radius of gyration estimated is 80 Å(Intensities (I(q)) are plotted in logarithmic scale while momentum transfer ($q^2$) is in $Å^{-1}$). Bottom left panel, Pair distribution function (P(r)) (P(r), probability distribution; r, interatomic vector length in Å). The bottom right panel, the fitting of the scattering curve (Intensities (I(q)) are plotted in logarithmic scale while momentum transfer (q) is in $Å^{-1}$).

the intensity of radiation scattered at zero angle. This value cannot be directly measured, however it can be extrapolated [Jacques and Trewhella, 2010]. I(0) is a measure of the number of scattering particles per unit volume and the particle volume. In other words, it is directly related to the concentration and the number of atoms in the particle [Jacques and Trewhella, 2010].

By plotting the data as ln(I(q)) versus $q^2$, the slope of linear fit of the data points is equal to $-R_g^2/3$ and thus $R_g$ is promptly calculated. I(0) can be extrapolated by looking at the y–intercept [Jacques and Trewhella, 2010]. In this study Guinier approximation was used to estimate $R_g$ and I(0) from the merged curve (Figure 6.2). The analysis was carried out in GNOM [Svergun, 1992]. The $R_g$ and I(0) estimated for *Pab*MCM were 80 Å and 0.27, respectively.

**Pair distance distribution function** The pair distance distribution function (P(r)) describes the probable frequency of the interatomic vector lengths (r) within a protein. This information can be obtained by Fourier transform the scattering profile [Jacques and Trewhella, 2010]. The P(r) profile, also known as radial Patterson function, is sensitive to the shape, symmetries and volume occupied by a protein or protein complex [Jacques and Trewhella, 2010].

The P(r) calculations depend upon indirect Fourier transform methods [Glatter, 1977]. The P(r) is estimated upon assumptions that P(r)=0 at zero and $P(D_{max})\geq 0$ at the $D_{max}$, with $D_{max}$ being the longest vector within the particle. As a result, $D_{max}$ is a model parameter in the interpretation of scattering data [Jacques and Trewhella, 2010]. Additionally, the P(r) function give also a more precise estimation of $R_g$ and I(0) since the entire scattering profile is used to calculate them [Glatter, 1977].

The P(r) distribution function, was obtained by indirect Fourier transformation of the scattering data by using the program GNOM [Svergun, 1992] (Figure 6.2). $R_g$, I(0) and $D_{max}$ estimated were 84 Å, 0.28, 280 Å, respectively.

Similar results have been reported after measuring the solution structure of the full-length *Mth*MCM by SANS [Krueger et al., 2014]. Data modelling of the scattering curve of the full-length *Mth*MCM suggested double hexameric structure [Krueger

et al., 2014].

## 6.3 Modelling the structure of the full-length *Pab*MCM in solution

The program DAMMIN was used to model the merged scattering curve of the full-length *Pab*MCM. DAMMIN implements a method, simulated annealing procedure, to restore *ab initio* low resolution shape of randomly oriented particles in solution from its SAXS experimental data [Svergun, 1999]. More specifically, DAMMIN calculate a single phase dummy atom model whose theoretical scattering fits the experimental data [Svergun, 1999]. A number of independent runs were performed with data to 55 Å. This resolution cut–off was chosen for several reasons, among which, the possibility of that the buffer may contribute to artefacts when higher q are used for modelling. The models were aligned, averaged, and then filtered by DAMAVER [Volkov and Svergun, 2003] to obtain a 'most probable' model for *Pab*MCM. Fitting of the theoretical scattering curve of the final model with experimental data was achieved using CRYSOL [Svergun et al., 1995]. A shown in Figure 6.3A (DAMFILT), the experimental filetered model of the in–solution structure of the full-length *Pab*MCM resembles a featureless double ring. The curve of the experimental filtered model of *Pab*MCM fits quite well the experimental data in the Guinier region, thus indicating that Rg and $D_{max}$ were estimated correctly. The unfiltered model does not fit the experimental curve as good as the filtered one, especially in the Guinier region. This might be explaned by the fact that filtering removes low occupancy and loosely connected atoms, hence generating a more compact model from the experimental data (Figure 6.3C).

In order to address whether the experimental curve and model of *Pab*MCM determinated by SAXS could be fitted with the theoretical curve calculated from the double octameric ring of *Pab*MCM, a double octameric ring model (Figure 6.3A, MODEL) was generated from a single octameric EM model (Figure 5.5). This model was generated manually in Chimera [Pettersen et al., 2004]. As shown in Figure 6.3B, the double octameric ring model fits quite nicely with the experimen-

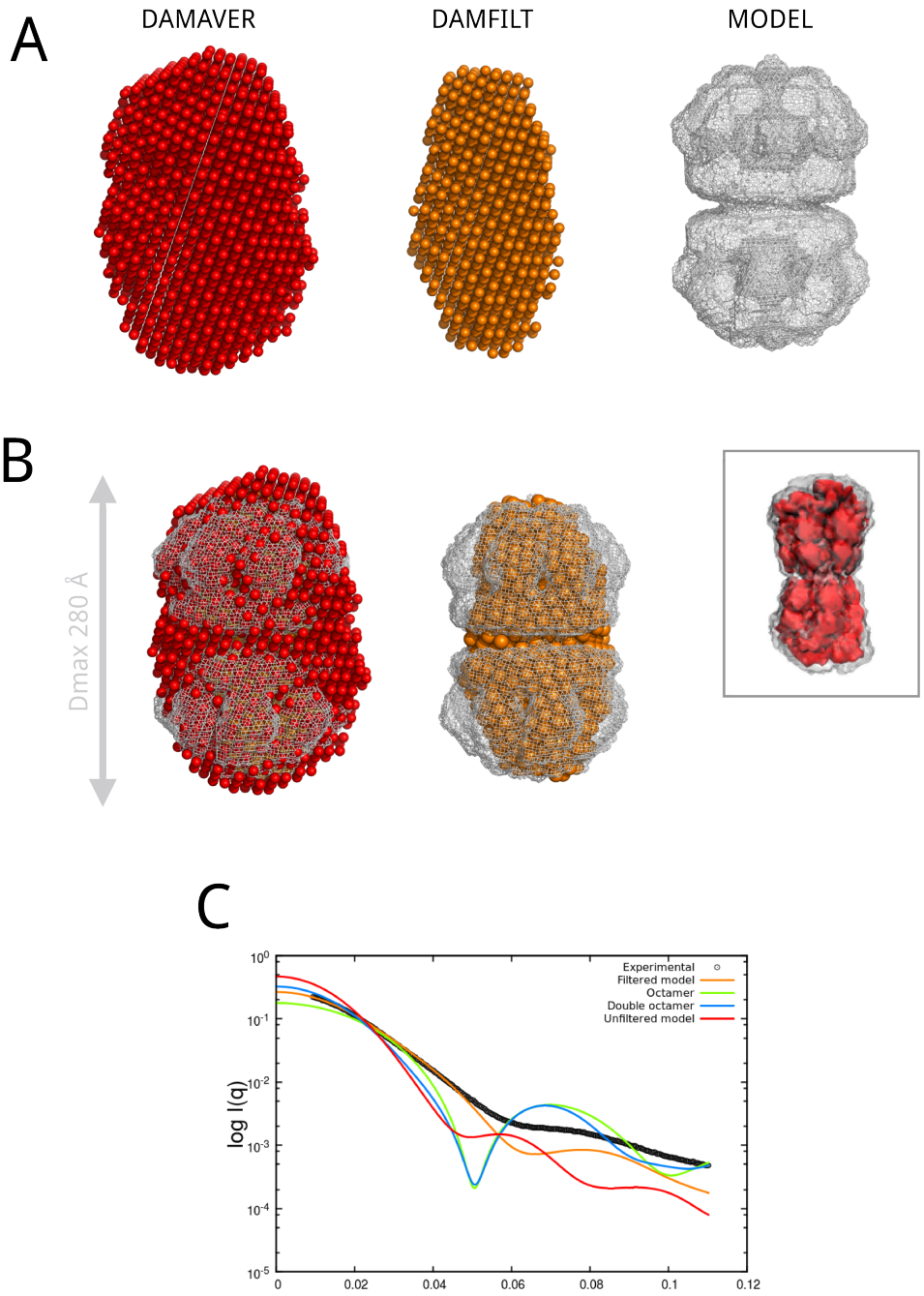



**Figure 6.3: SAXS *ab initio* model of the *Pab*MCM.**. (A) DAMAVER, average of 50 models. DAMFILT, the averaged model at a given cut–off volume. MODEL, EM model built based on the 3D–EM map of the octameric *Pab*MCM. (B) All models in A are superimposed. The $D_{max}$ is indicated. In the grey box, SANS *Mth*MCM model for comparison (not in scale) [Krueger et al., 2014]. (B) Fitting of the theoretical scattering curves, of the models in A, to the experimental data(Intensities (I(q)) are plotted in logarithmic scale while momentum transfer (q) is in $Å^{-1}$).

tal filtered SAXS model. Interestingly, a similar model was recently proposed for the full-length *Mth*MCM complex investigated by SANS [Krueger et al., 2014] (Figure 6.3B, boxed model). The theoretical curve of the double octameric ring model of *Pab*MCM calculated by CRYSOL, fits to the experimental SAXS curve of *Pab*MCM quite nicely in the Guinier region (Figure 6.3C). However, the fitting becomes less good toward higher q. This could be explained by the fact that the sample is not monodisperce (as shown in chapter 5) and hence the experimental curve is an average of different assemblies present in the sample. It is not possible to asses the polydispersivity of *Pab*MCM and give a precise estimation of the frequency of the double rings since the double rings being twice as much bigger than the single rings will scatter more than the single rings hence giving a curve whose overall average is weighted more toward the double rings. Therefore, it is not possible to provide and accurate estimation of the frequency of both assemblies. Although, these preliminary data support a model in which, the full-length *Pab*MCM adopts a double ring assembly in solution, EM data have shown that *Pab*MCM is a mixture of at least four assemblies (chapter 5) and hence more biophysical characterisation (Dynamic light scattering, Size-exclusion chromatography and multiangle scattering) needs to be carried out to better asses the polydispersivity of *Pab*MCM in solution.

## 6.4 Concluding remarks

Small–angle scattering (SAS) is a powerful tool used for investigating the structure of biological samples in solution [Feigin et al., 1987]. SAXS on the full-length *Pab*MCM suggested that *Pab*MCM assembles in a big protein complex with a $R_g$ of 84 Å. By P(r) analysis, the largest distance ($D_{max}$) within the complex was measured to be 280 Å. Modelling of the SAXS data revealed that *Pab*MCM in solution assemble in a big protein complex comparable in size to a double ring model. These data support previous observation by EM studies (section 5). Previous SANS experiments with related proteins reported similar value [Krueger et al., 2014].

# Chapter 7

# Results and Discussion: structural insights in the Okazaki fragments maturation protein complex, the "Okazakisome"

## 7.1 Introduction

The crenarchaeon *Sulfolobus solfataricus* possesses a simplified toolset for DNA replication compared to Eukaryotes, and is therefore used as a model system for the study of replication proteins [Dionne et al., 2003]. Interestingly, *S. solfataricus* has a subset of the eukaryotic Okazaki fragment maturation factors, among which there are a heterotrimeric DNA sliding clamp, the proliferating cell nuclear antigen (*Sso*PCNA), the DNA polymerase B1 (*Sso*PolB1), the flap endonuclease (*Sso*Fen1) and the ATP-dependent DNA ligase I (*Sso*LigI). *Sso*PCNA has been demonstrated to function as a scaffold with each subunit having a specific binding affinity for each of the factors involved in Okazaki fragment maturation [Dionne et al., 2003]. Moreover, Beattie and Bell [2012] demonstrated that the most efficient coupling of activities occurs when a single *Sso*PCNA ring organises *Sso*PolB1, *Sso*Fen1 and *Sso* LigI into a complex. Thus, these proteins are necessary and sufficient for concerted

DNA synthesis on the lagging strand. Here we show the 3D reconstruction of the "Okazakisome", in other words *Sso*PCNA in complex with the Okazaki fragment maturation proteins *Sso*PolB1, *Sso*LigI and *Sso*Fen1.

## 7.2 Protein expression and purification of the recombinant Okazaki fragment maturation proteins

Recombinant *Sso*PCNA fusion protein, *Sso*PolB1, *Sso*Fen1 or *Sso*LigI were expressed in Rosetta$^{TM}$(DE3) pLysS host cells harbouring pET33b–PCNA123-6His, pET33b–Polb1, pET33b–Fen1 or pET30a–LigI-6His constructs. Recombinant protein expression was checked by SDS–PAGE (Figure 7.1 A). As shown, proteins were highly expressed. It is important to point out that the construct pET33b–PCNA123-6His expresses *Sso*PCNA as fusion protein in which the three subunits, composing the ring, are fused together by a polyglycine stretch. The recombinant fusion protein is functional as shown in Dionne et al. [2003].

*S. solfataricus* is an acidophilic and thermophilic archaeon whose optimal growth conditions occur at pH 2.0–3.0 and temperatures of 75–80°C. As already shown for other archaeal proteins, one of the most impressive proprieties of these proteins is their thermostability. In order to take advantage of this property, after cell lysis achieved by sonication (*Sso*PCNA123, *Sso*PolB1, *Sso*Fen1) and cell–grinding in liquid nitrogen (*Sso*LigI), the soluble fraction was heat–treated (Figure 7.1 B).

Heat-resistant soluble fractions were taken forward for further purification steps (purification protocols of *Sso*PCNA123, *Sso*PolB1, *Sso*Fen1 and *Sso*LigI are listed in Table 3.11). *Sso*PCNA was purified at final concentration of 2.2 mg/ml. (Figure 7.2). The isolated protein was quite pure with relatively fewer smaller bands appearing on SDS–PAGE. *Sso*PolB1 was isolated at final concentration of 2.7 mg/ml (Figure 7.3, Figure 7.4, Figure 7.5). The protein in this case was essentially pure and homogeneous. *Sso*Fen1 was purified at final concentration of 2.1 mg/ml (Figure 7.6, Figure 7.7. *Sso*LigI was isolated final concentration of 1.57 mg/ml (Figure 7.8,

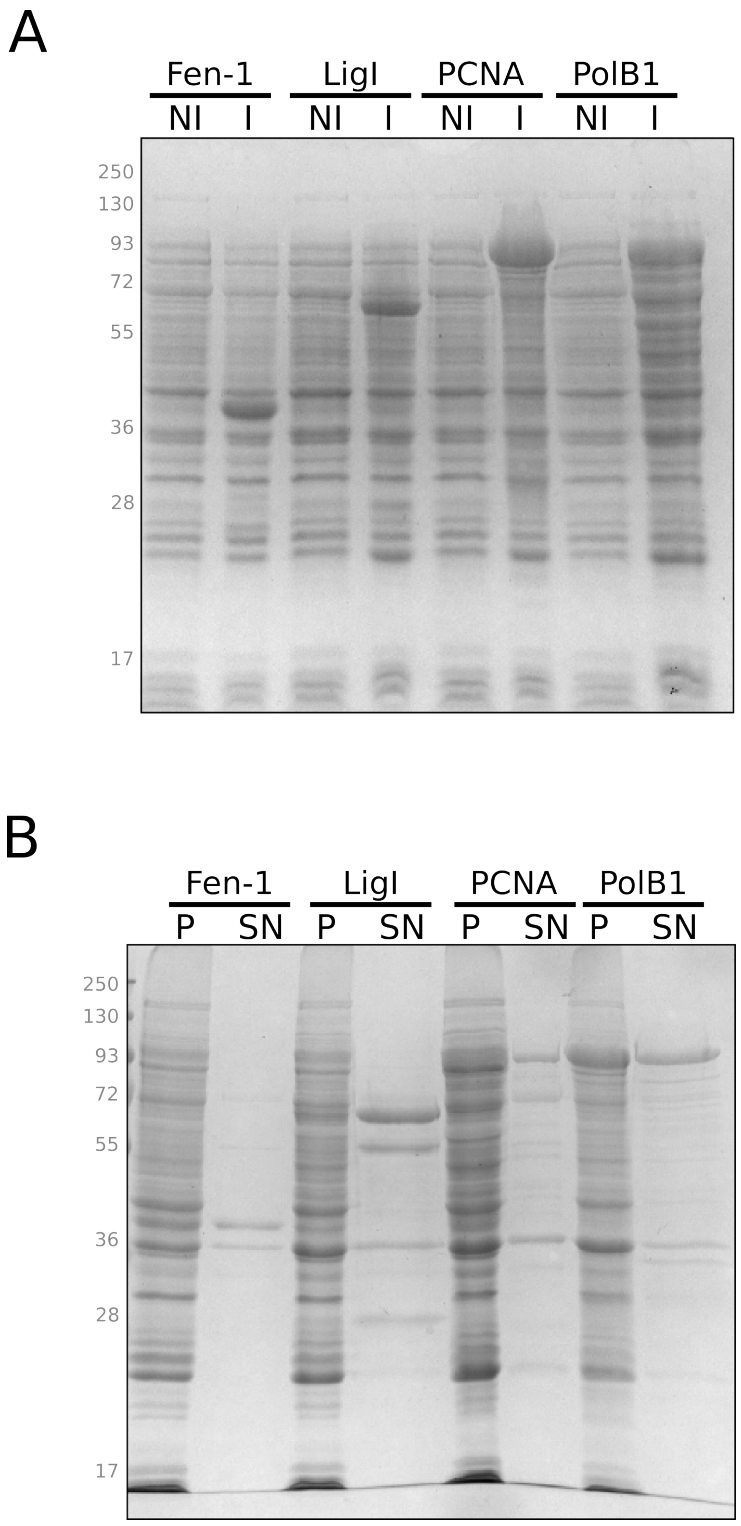



**Figure 7.1: Protein expression and heat-denaturation of the Okazaki fragment maturation proteins.** Recombinant protein expression was carried out in Rosetta$^{TM}$(DE3) pLysS host cells. Cells were induced with 1 mM IPTG when the $OD_{600}$ was 0.6 – 0.8. (A) SDS-PAGE analysis of protein over-expression in crude extracts. NI, not induced host cells. I, Induced host cells. (B) SDS-PAGE analysis of soluble (SN) and insoluble (p) fractions after heat-denaturation. Heat-denaturation was performed at 65°C for PolB1 whilst Fen1, Lig1 and PCNA total lysates were treated 75°C. (A, B) Visualisation by Coomassie stain, molecular weight markers are labelled on the left, sizes are in kDa.

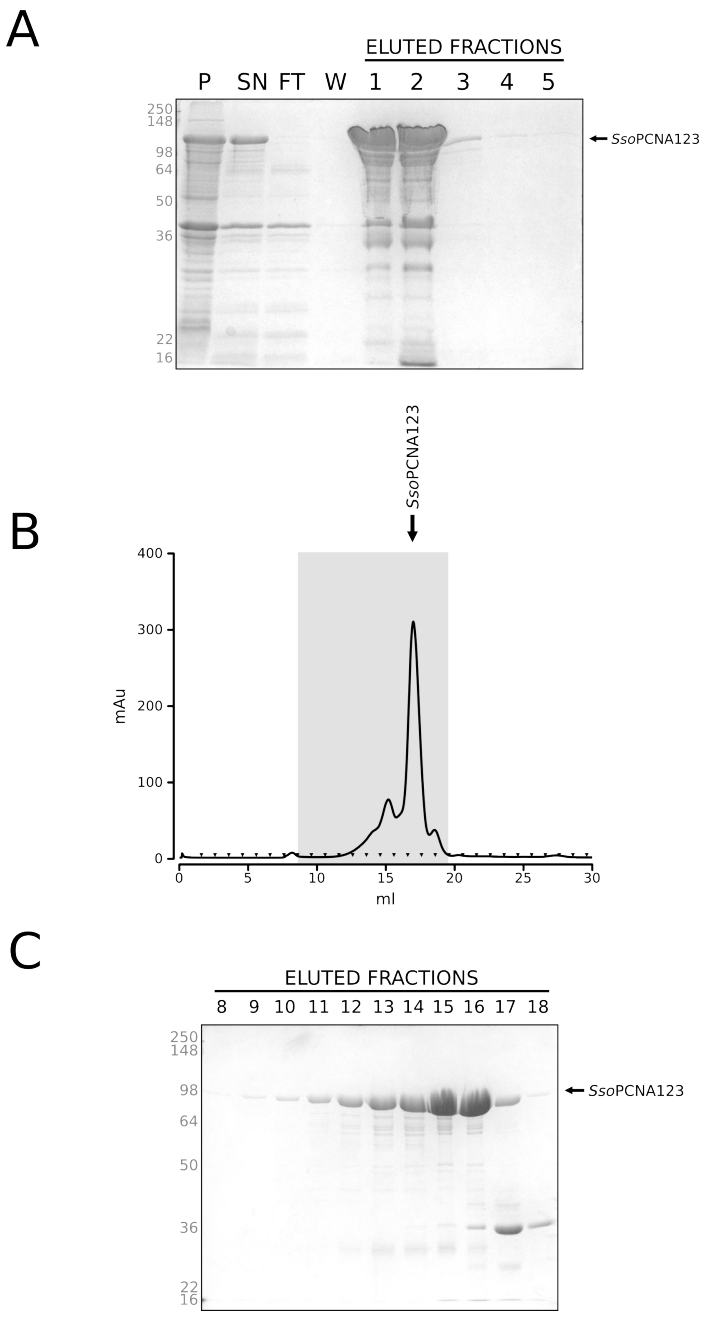



**Figure 7.2: Purification of the fused heterotrimeric proliferative cell nuclear antigen *Sso*PCNA123.** (A) SDS–PAGE analysis of nickel affinity purification. Chromatography was performed by loading the supernatant (SN) onto a 1 ml HisTrap$^{TM}$(GE) pre-packaged column with a syringe. The column was washed with 30 ml of 20 mM HEPES pH 8.0, 300 mM NaCl (W). Protein was eluted with buffer 20 mM HEPES pH 8.0, 300 mM NaCl, 500 mM imidazole. Five fractions of 1 ml each were collected (1-5). (P) Pellet. (FT) Flow-through collected during loading. (B) Size-exclusion chromatography. Fraction 1,2 and 3 were pooled together and concentrated up to 500 $\mu$l and then loaded onto a Superose6$^{TM}$10/300 GL. Chromatography was performed at 0.3 mlmin flow rate in 20 mM HEPES pH 8.0, 300 mM NaCl. Light grey shading indicates the part of the chromatogram being analysed by SDS–PAGE. (C) SDS–PAGE analysis of the fractions collected after size-exclusion chromatography. (8–18) Eluted fractions. Bands were visualised by Coomassie stain. On the left–side markers in kDa.

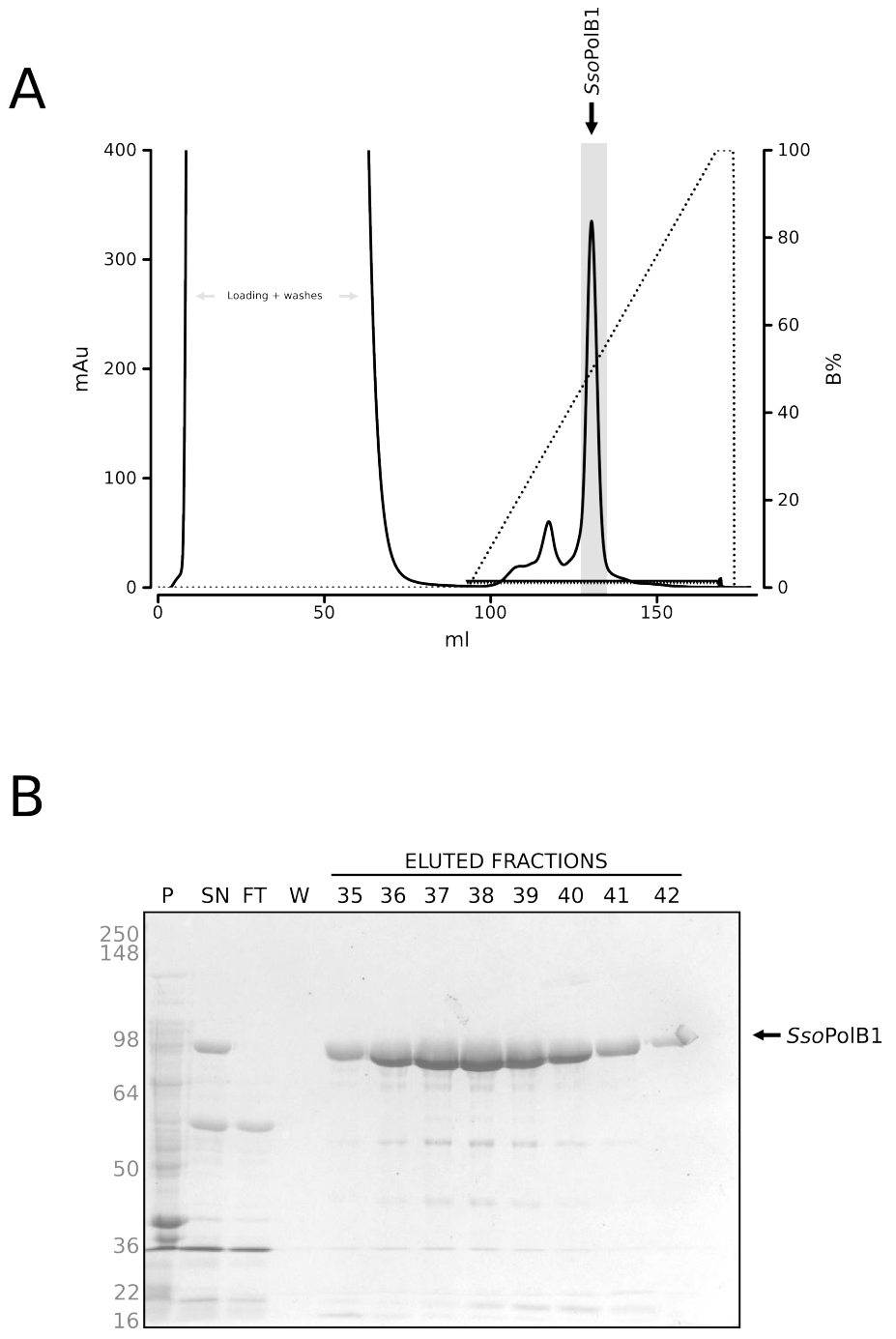



**Figure 7.3: Purification of the polymerase *Sso*PolB1 (First step).** (A) Heparin affinity chromatography. Chromatography was performed with a 1 ml HiTrap$^{TM}$Heparin (GE) pre-packaged column, pre-equilibrated with 10 mM HEPES pH 7.5, 100 mM NaCl, 1 mM DTT. Sample was eluted with a 15 ml linear gradient of buffer B (20 mM HEPES pH 7.5, 1000 mM NaCl, 1 mM DTT). Collected fractions were analysed by SDS–PAGE. (B) SDS–PAGE analysis of the fractions collected during heparin affinity chromatography. (P) Pellet. (SN) Supernatant loaded onto the 1 ml HiTrap$^{TM}$Heparin (GE) pre-packaged column. (FT, W) Flow-trough and washes collected during loading (Loading + washes). (35–42) Fractions collected during the chromatography (light grey shading in panel A). Bands were visualised by Coomassie stain. On the left–side markers in kDa.

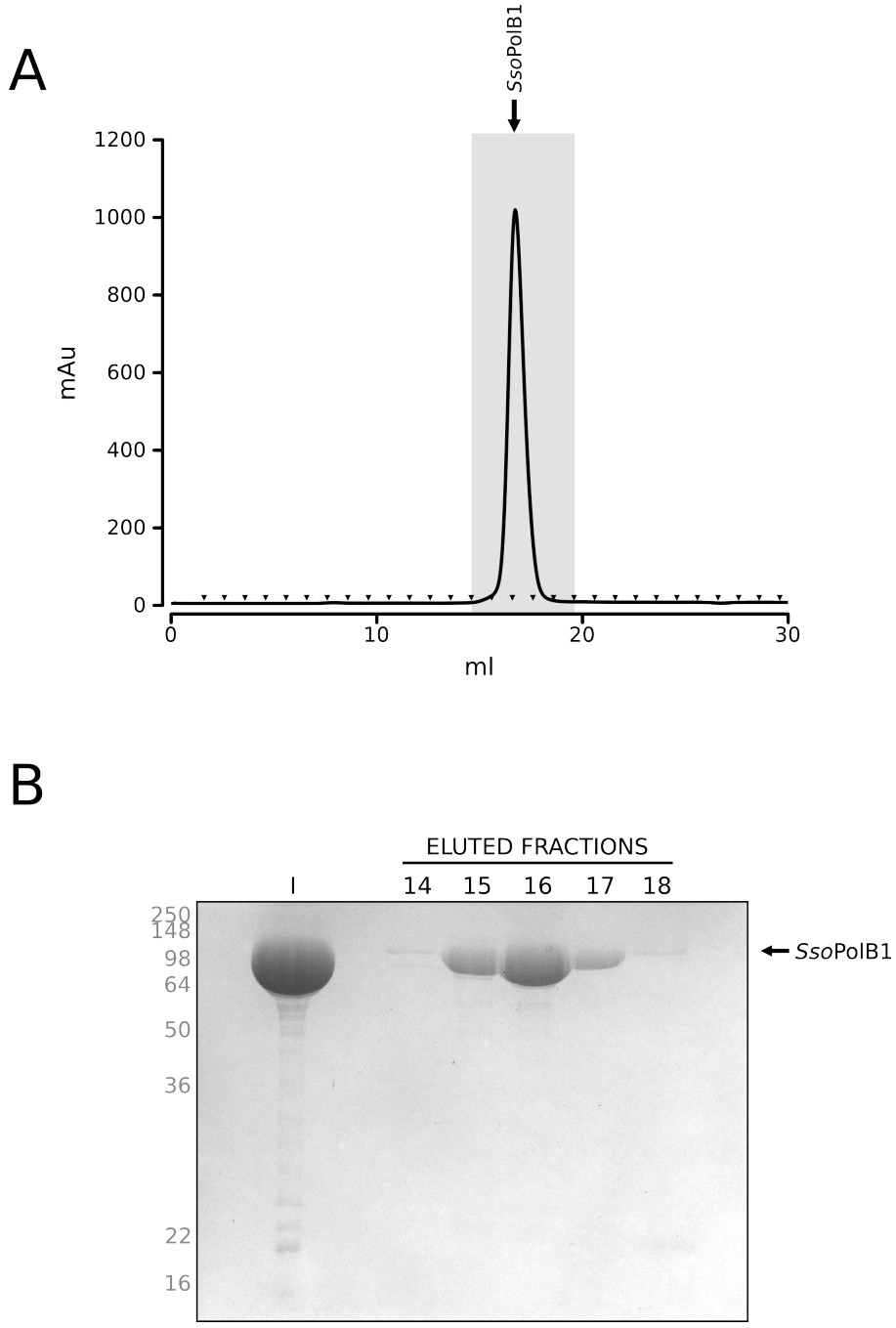



**Figure 7.4: Purification of the polymerase *Sso*PolB1 (Second step).** (A) Size–exclusion chromatography. Fractions from 35 to 42 from the first step purification (Figure 7.3) were pooled and concentrated to a final volume of 500 $\mu$l which was then loaded onto a Superose6$^{TM}$10/300 GL column pre-equilibrated with 20 mM HEPES pH 7.5, 500 mM NaCl, 1 mM DTT. Chromatography was performed at 0.3 ml/min flow rate. Collected fractions were analysed by SDS–PAGE. (B) SDS–PAGE analysis of the fractions collected during size–exclusion chromatography. (I) Concentrate sample loaded onto the column. (14–18) Fractions collected during size-exclusion chromatography (light grey shading in panel A). Bands were visualised by Comassie stain. On the left–side markers in kDa.

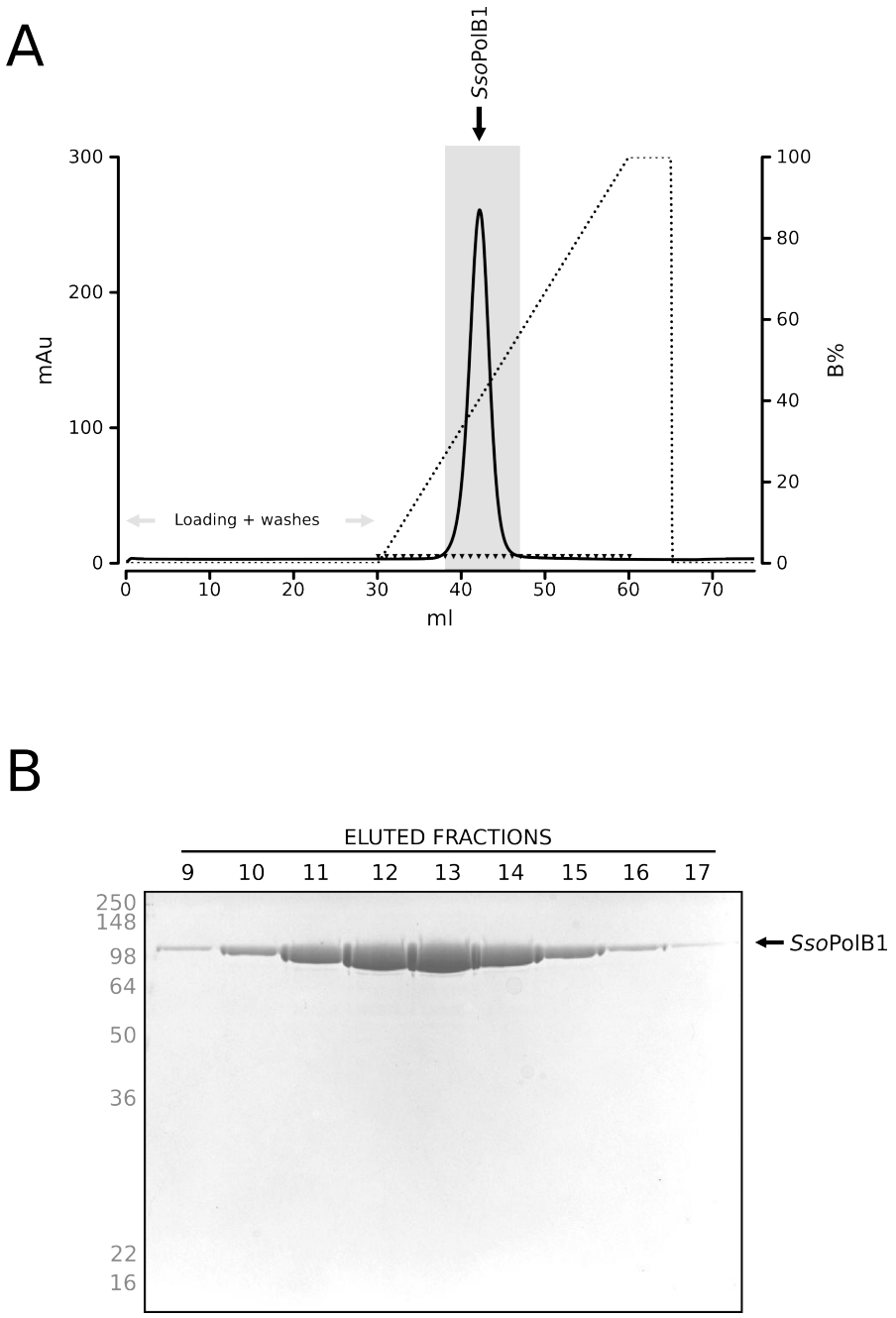



**Figure 7.5: Purification of the polymerase *Sso*PolB1 (Third step).** (A) Strong anion exchange chromatography. Fractions from 15 to 17 from the second step purification (Figure 7.4) were pooled and concentrated to a final volume of 1 ml which was then loaded onto a HiTrap$^{TM}$S column pre-equilibrated with 20 mM HEPES pH 7.5, 100 mM NaCl, 1 mM DTT. Sample was eluted with a 15 ml linear gradient of buffer B (20 mM HEPES pH 7.5, 1000 mM NaCl, 1 mM DTT). Collected fractions were analysed by SDS–PAGE. (B) SDS–PAGE analysis of the fractions collected during strong anion exchange chromatography. (9–17) Fractions collected during size-exclusion chromatography (light grey shading in panel A). Bands were visualised by Comassie stain. On the left–side markers in kDa.

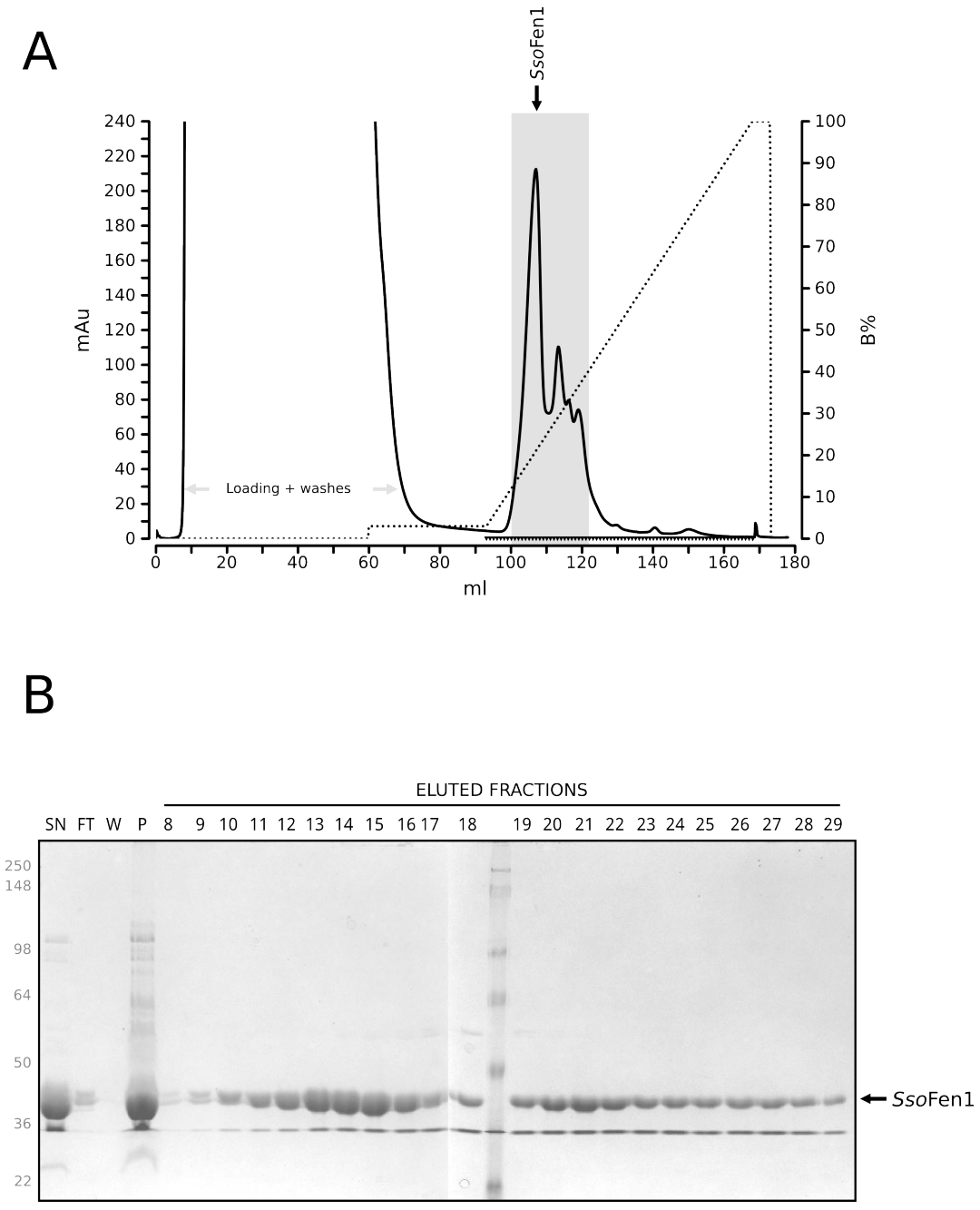



**Figure 7.6: Purification of the flap-endonuclease *Sso*Fen1 (First step).** (A) Heparin affinity chromatography. Chromatography was performed with a 1 ml HiTrap$^{TM}$Heparin (GE) pre-packaged column, pre-equilibrated with 20 mM MES pH 6.0, 30 mM NaCl, 1 mM EDTA, 0.5 mM DTT. Sample was eluted with a 15 ml linear gradient of buffer B (20 mM MES pH 6.0, 1000 mM NaCl, 1 mM EDTA, 0.5 mM DTT). Collected fractions were analysed by SDS–PAGE. (B) SDS–PAGE analysis of the fractions collected during heparin affinity chromatography. (SN) Supernatant loaded onto the 1 ml HiTrap$^{TM}$Heparin (GE) pre-packaged column. (FT, W) Flow-trough and washes collected during loading (Loading + washes). (P) Pellet. (8–29) Fractions collected during the chromatography (light grey shading in panel A). Bands were visualised by Comassie stain. On the left–side markers in kDa.

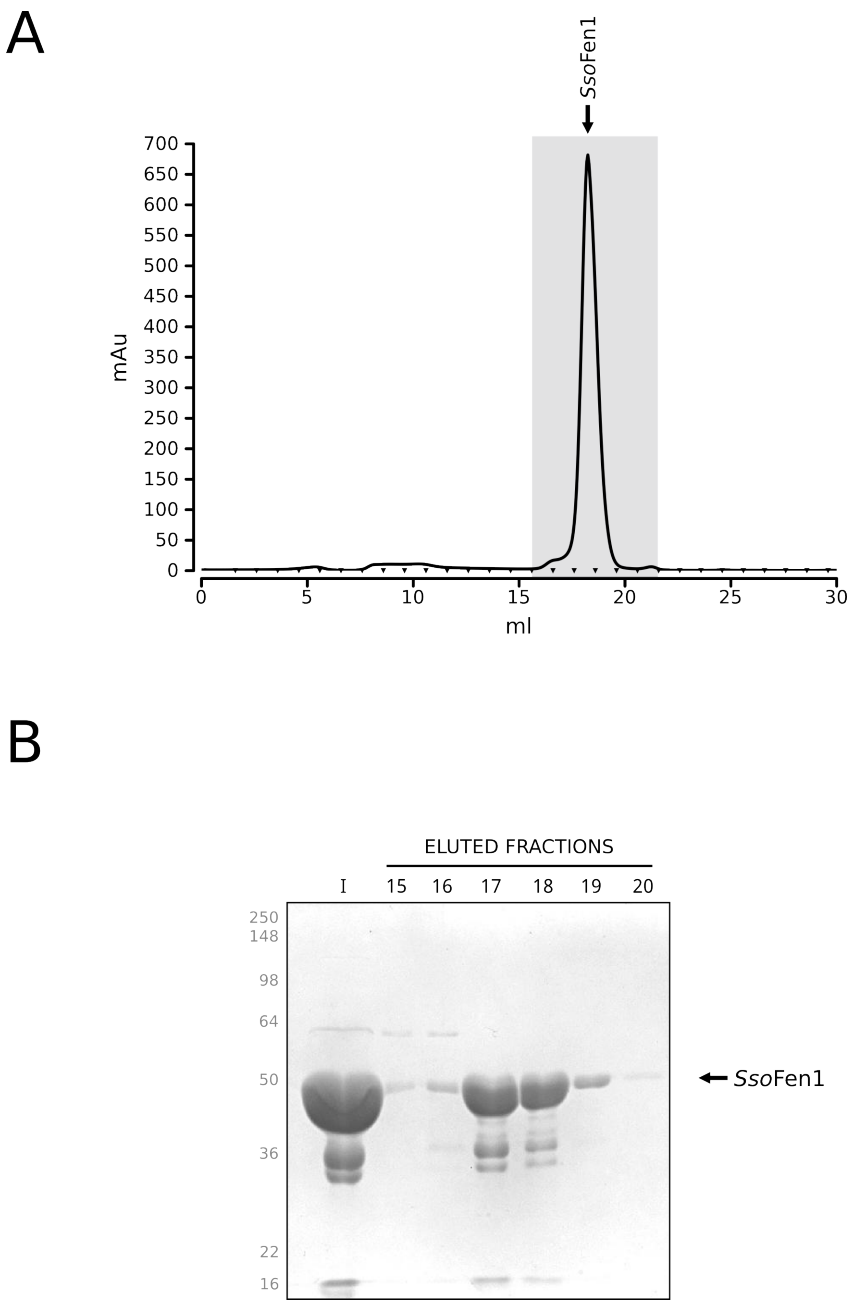



**Figure 7.7: Purification of the polymerase *Sso*Fen1 (Second step).** (A) Size–exclusion chromatography. Fractions from 9 to 18 from the first step purification (Figure 7.6) were pooled and concentrated to a final volume of 500 $\mu$l which was then loaded onto a Superose6$^{TM}$10/300 GL column pre-equilibrated with 20 mM MES pH 6.0, 150 mM NaCl, 1 mM EDTA, 0.5 mM DTT. Chromatography was performed at 0.3 ml/min flow rate. Collected fractions were analysed by SDS–PAGE. (B) SDS–PAGE analysis of the fractions collected during size–exclusion chromatography. (I) Concentrate sample loaded onto the column. (15–20) Fractions collected during size-exclusion chromatography (light grey shading in panel A). Bands were visualised by Comassie stain. On the left–side markers in kDa.

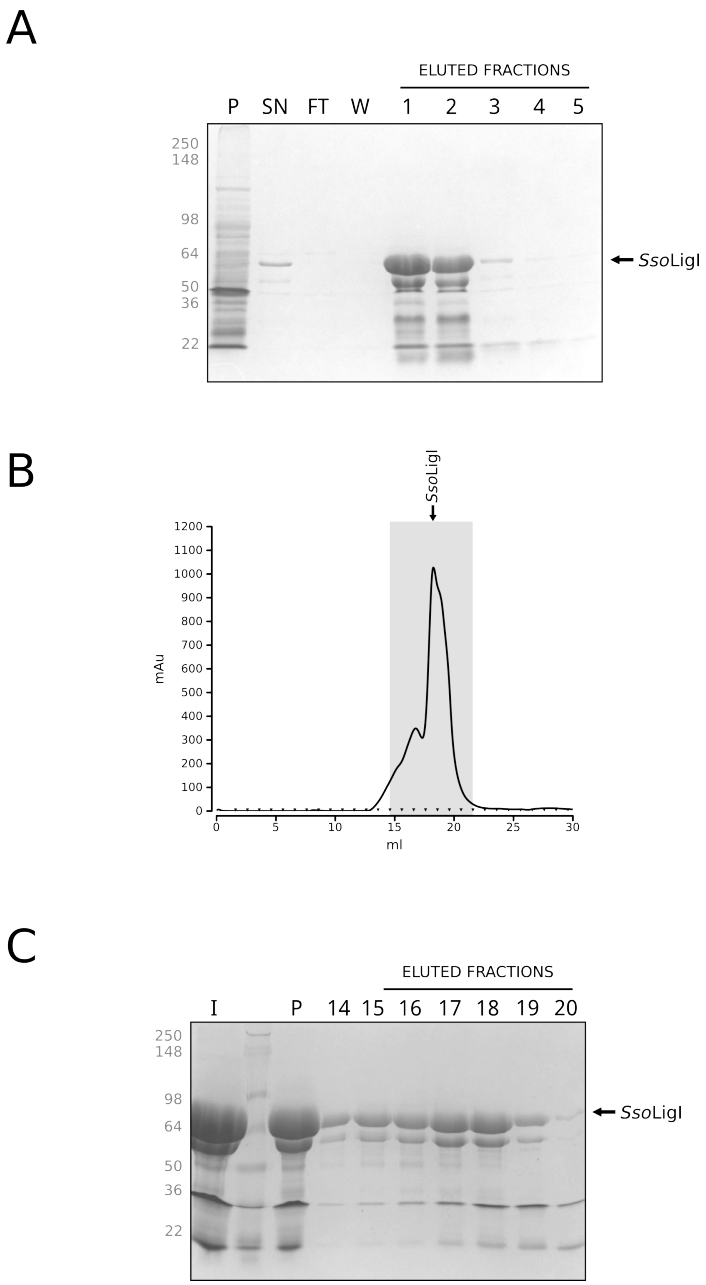



**Figure 7.8: Purification of the ligase *Sso*LigI (First and second step).** (A) SDS–PAGE analysis of nichel affinity purification (first step). Chromatography was performed by loading the supernatant (SN) onto a 1 ml HisTrap$^{TM}$(GE) pre-packaged column, pre-equilibrated with 10 mM HEPS pH 8.0, 300 mM NaCl, with a syringe. Bounded protein was washed with 30 ml of 20 mM HEPES pH 8.0, 300 mM NaCl (W). Protein was eluted with buffer 20 mM HEPES pH 8.0, 300 mM NaCl, 500 mM imidazole. Five fractions, of 1 ml each, were collected (1-5). (P) Pellet. (FT) Flow-trough collected during loading. (B) Size-exclusion chromatography (second step). Fraction 1,2 and 3 were pooled together and concentrated up to 500 $\mu$l and then loaded onto a Superose6$^{TM}$10/300 GL. Chromatography was performed at 0.3 mlmin flow rate in 20 mM HEPES pH 8.0, 150 mM NaCl, 14 mM $\beta$-mercaptoethanol. Light grey shading indicates the part of the chromatogram being analysed by SDS–PAGE. (C) SDS–PAGE analysis of the fractions collected after size-exclusion chromatography. (I) Concentrate sample loaded onto the column. (P) Pellet, protein precipitated during concentration by ultra-filtration. (14–20) Eluted fractions. Bands were visualised by Comassie stain. On the left–side markers in kDa.

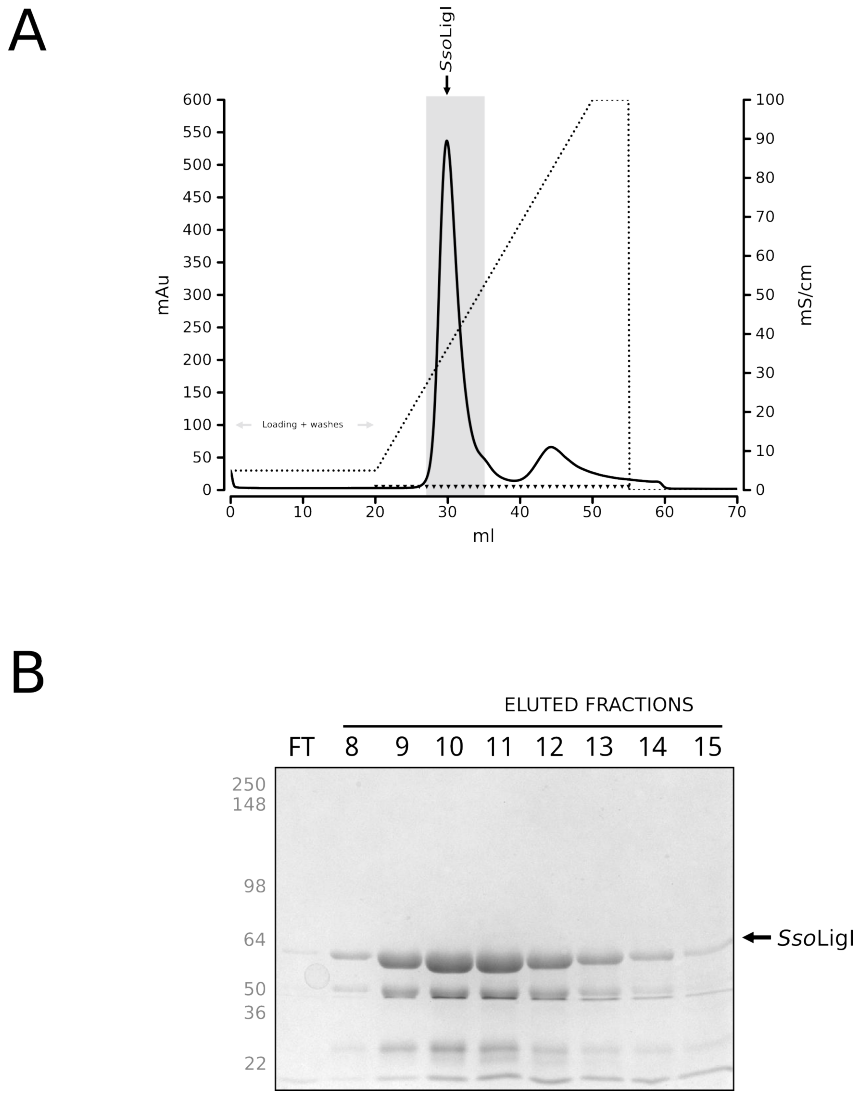



**Figure 7.9: Purification of the ligase *Sso*LigI (Third step).** (A) Strong cataion exchange chromatography. Fractions from 17 to 19 from the second step purification (Figure 7.8) were pooled and concentrated to a final volume of 1 ml which was then loaded onto a HiTrap$^{TM}$Q column pre-equilibrated with 20 mM HEPES pH 8.0, 75 mM NaCl, 14 mM $\beta$-mercatpoethanol. Sample was eluted with a 15 ml linear gradient of buffer B (20 mM HEPES pH 8.0, 1000 mM NaCl, 14 mM $\beta$-mercatpoethanol). Collected fractions were analysed by SDS–PAGE. (B) SDS–PAGE analysis of the fractions collected during strong cataion exchange chromatography. (FT) Flow-trough collected during loading. (8–15) Fractions collected during size-exclusion chromatography (light grey shading in panel A). Bands were visualised by Comassie stain. On the left–side markers in kDa.

Figure 7.9.

## 7.3 Reconstitution of the 'Okazakisome' complex on DNA

The predicted *Sso*PCNA123–*Sso*PolB1–*Sso*Fen1–*Sso*LigI•DNA complex [herein Okazakisome] was reconstructed from purified recombinant proteins onto a DNA structure mimicking two adjacent Okazaki fragments. The DNA oligonucleotides were designed in order to feature a mismatch and a 5'–flap when annealed (Figure 7.10 A) [Tsutakawa et al., 2011]. DNA mimicking two adjacent Okazaki was prepared at final concentration of 30 $\mu$M as follow:

**Reaction mix:**

| **Component** | **Volume** |
|---:|---|
| Upstream (300 $\mu$M) | 1 $\mu$l |
| Downstream (300 $\mu$M) | 1 $\mu$l |
| Template (300 $\mu$M) | 1 $\mu$l |
| 10X Annealing buffer | 1 $\mu$l |
| Nuclease–free $H_2O$ | 9 $\mu$l |
| **Total volume** | **10 $\mu$l** |

Oligos were mixed together and then incubated on a thermocycler for 5 minutes at 95°C; 5 minutes at 90°C; 5 minutes at 85°C; 5 minutes at 80°C and then let the mixture cool down to room temperature for an hour. Annealing was checked by native 8% acrylamide gels (data not shown).

The Okazakisome was formed by incubating *Sso*PCNA123 protein with 2x excess DNA oligonucleotide in 10 mM HEPES pH 8.0, 150 mM NaCl, 5 mM $MgCl_2$ at 50°C for 50 minutes. After then, equimolar amount of *Sso*PolB1, *Sso*Lig1 and *Sso*Fen1 were added into the mix and incubated for more 30 minutes. Reaction mix was composed as follow:

| **Reaction Mix:** | |
|---|---|
| **Component** | **Volume** |
| *Sso*PCNA123 (25 $\mu$M) | 10 $\mu$l |
| *Sso*PolB1 (25 $\mu$M) | 10 $\mu$l |
| *Sso*LigI (25 $\mu$M) | 10 $\mu$l |
| *Sso*Fen1 (50 $\mu$M) | 5 $\mu$l |
| DNA (30 $\mu$M) | 20 $\mu$l |
| 10X buffer | 20 $\mu$l |
| Nuclease–free $H_2O$ | 125 $\mu$l |
| **Total volume** | **200 $\mu$l** |

## 7.4 Purification of the "Okazakisome" complex

The reconstructed Okazakisome was then purified using GraFix method, a sample preparation method for EM single-particles analysis, in which a glycerol gradient is coupled with a glutaraldehyde gradient [Kastner et al., 2007]. This method has been shown to be very effective when working with highly dynamic as well as fragile protein complex [Wang et al., 2009; Kastner et al., 2007; Herzog et al., 2009]. GraFix gradient was fractionated manually from the top of the tube. Collected fractions were analysed by NuPAGE (Figure 7.10 B). As shows, there is a band-shift, toward the higher molecular weight, from the fraction $5^{th}$ to the fraction $10^{th}$, which likely indicates either the presence of different conformational state adopted during the Okazaki fragments maturation or the stabilization of the complex toward a more compact structure due to the increasing amount of cross-linker. The fraction $10^{th}$ appeared to be the pure and homogeneous in term of size. In order to establish whether or not in the fraction $10^{th}$ contained all four proteins, fractions were also probed for the presence of individual proteins by western blot (Figure 7.11 A). As shown, a positive signal was detected for *Sso*PCNA, *Sso*PolB1, *Sso*LigI and *Sso*Fen1 in the lane corresponding to the fraction $10^{th}$. Although the signal corresponding to *Sso*LigI and *Sso*Fen1 was quite narrow and localized, the signal for *Sso*PCNA123 and *Sso*PolB1 was wider but localized around the right molecular weight. Contaminations of *Sso*PCNA123 not bound were detected in the fraction $10^{th}$. However,

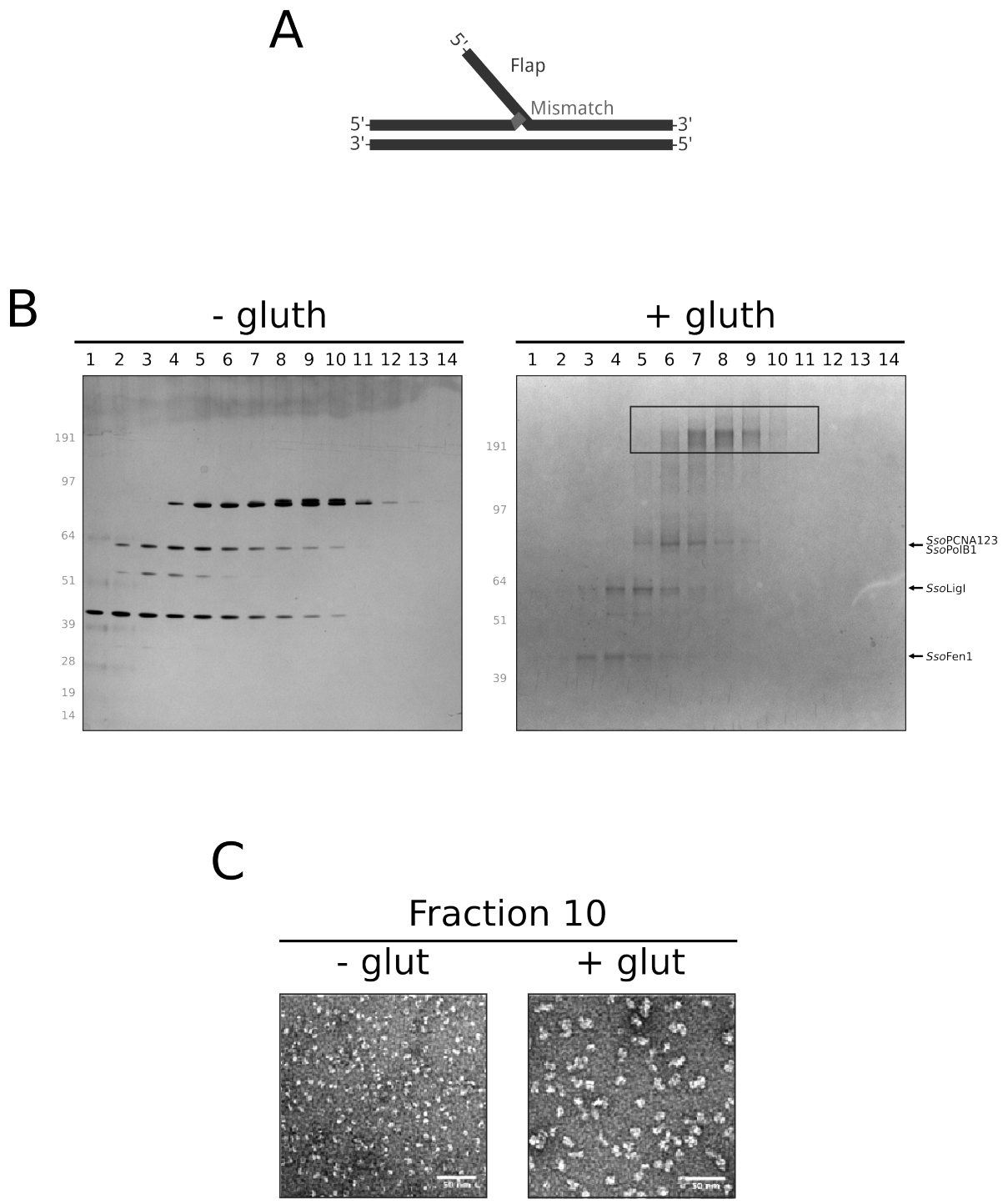



**Figure 7.10: Purification of the *Sso*PCNA–*Sso*PolB1–*Sso*Fen1–*Sso*LigI–DNA complex of *S. solfataricus*.** (A) Schematic of DNA structure used in the Grafix input. (B) Nu–PAGE analysis of the samples collected from GraFix gradients. The rectangle highlights the band-shift detected after running GraFix gradient. Bands are visualised by SimplyBlue$^{TM}$(G-250 Coomassie stain). Molecular markers are showed on the left side. (C) Two micrographs showing the fraction $10^{th}$ with and without glutaraldehyde. The complex fractionated from pycnic gradients not containing glutaraldehyde disassembles upon EM sample preparation.

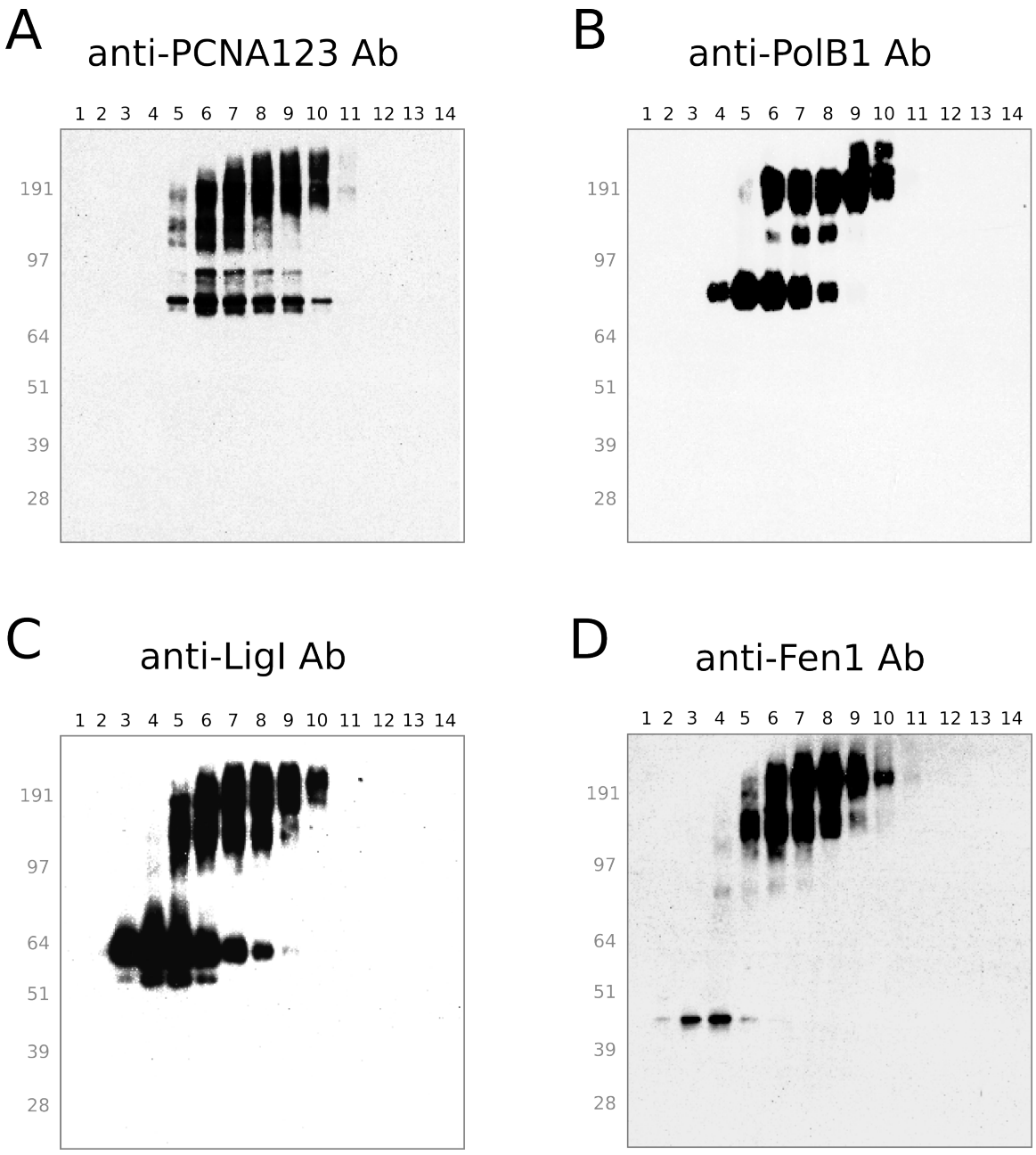



**Figure 7.11: Western blotting analysis of the fractions collected from GraFix.** (A) Western blot probed with anti-*Sso*PCNA123 antibody whereas (B) (C) (D) were probed with anti-*Sso*PCNA123, anti-*Sso*PolB1, anti-*Sso*LigI and anti-*Sso*Fen1, respectively.

*Sso*PCNA was not seen as contaminant in the raw micrographs, indicating that likely the amount of crosslinker was not saturating.

Collected fractions were also analysed by EM. Fractions isolated from the glycerol gradient without cross-linker were heterogeneous, indicating that the complex was fragile and not suitable for EM. This is also demonstrated comparing the fractions containing cross-linker, which were characterized by larger discrete and very homogeneously dispersed particles (Figure 7.10 C) indicating therefore that the protein complex was likely to be stabilized. Accordingly to all these results the fraction $10^{th}$ was chosen for EM single-particles analysis.

## 7.5 Electron microscopy studies of 'Okazakisome'

To elucidate the three–dimensional architecture of the Okazakisome, electron microscopy coupled to single particle analysis experiments was performed. A typical micrograph is shown in Figure 7.12 A. Reference-free 2D analysis showed that single particles could be classified to calculate a gallery of class averages with consistent features (Figure 7.12 B). We boxed ∼15000 single particle images, and used IMAGIC-5 [van Heel et al., 1996] and EMAN [Ludtke et al., 1999] protocols to calculate and refine the 3D volume. Projections of the maps matched with 2D class-averages assigned the same Euler angles (Figure 7.12 C), highlighting the validity of the map. The *Sso*PCNA123–*Sso*PolB1–*Sso*Fen1–*Sso*LigI•DNA complex (Figure 7.12 D) exhibits features, compatible with previous electron microscopy studies of related assemblies (*Pfu*PCNA–*Pfu*Lig•DNA [Mayanagi et al., 2009] and *Sso*PCNA–*Pfu*PolI•DNA [Mayanagi et al., 2011]. *Sso*PCNA–*Sso*PolB1–*Sso*Fen1–*Sso*Lig•DNA has overall dimensions of 150Å x 160 Å x 95 Å. The volume of this 3D reconstruction is compatible with a molecular weight of ∼350 kDa, which is the expected molecular weight for the *Sso*PCNA–*Sso*PolB1–*Sso*Fen1–*Sso*Lig•DNA complex. The resolution of the map is 22Å, calculated at 0.5 Fourier Shell Correlation (Figure 7.12 C).

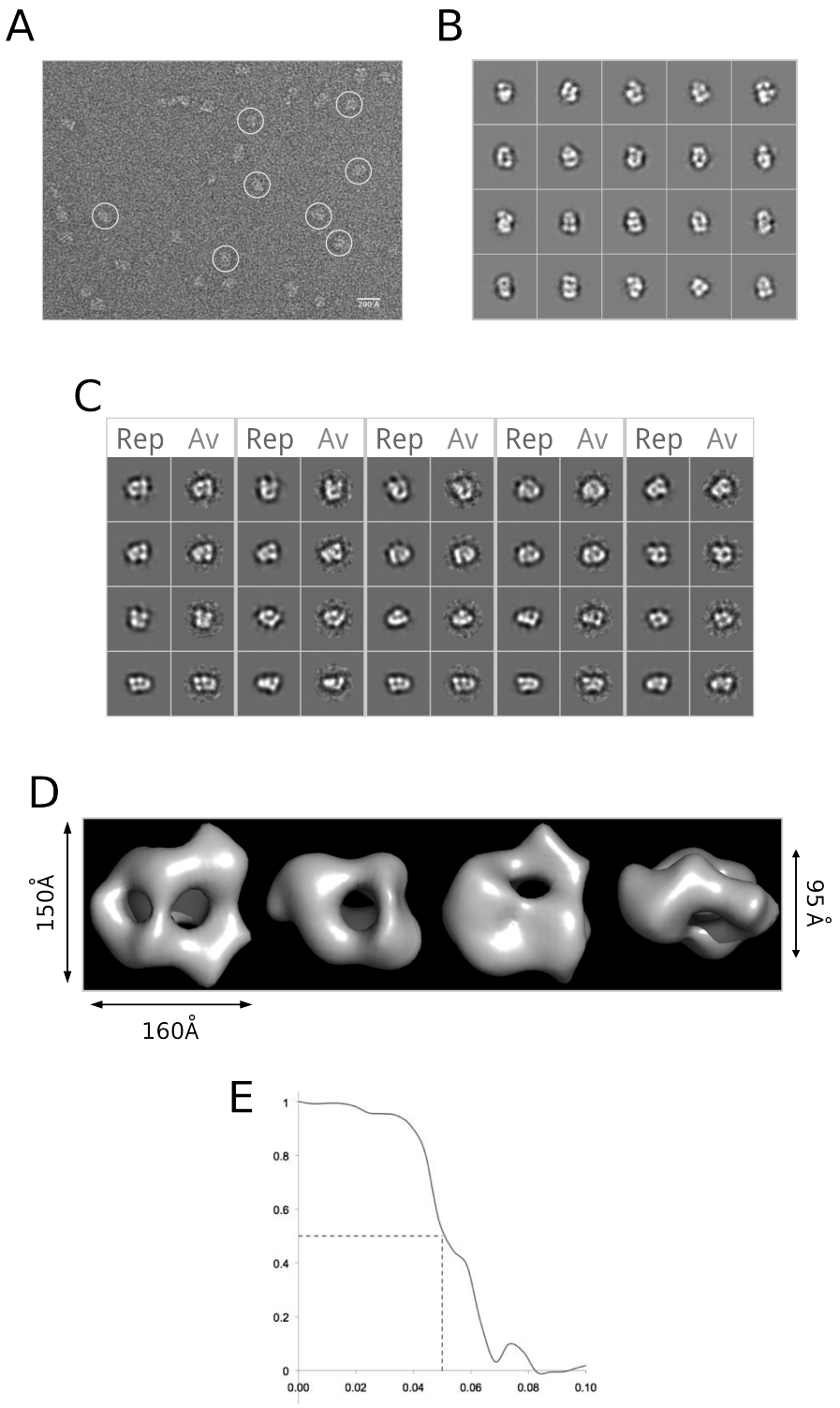



**Figure 7.12: Single particle analysis and 3D reconstruction of the Okazaki fragments maturation protein complex, "Okazakisome".** (A) Characteristic negatively stained micrograph. White circles are used to show single particles. (B) Class-average obtained by multi-reference alignment by classification with IMAGIC–5. Reference-free class averages show a three-layered architecture. (C) Rep, Reprojections; and Av, class–averages of the refined 3D reconstruction. (D) 3D model of the *Sso*PCNA–*Sso*PolB1–*Sso*Fen1–*Sso*Lig1•DNA complex. (E) Fourier shell correlation plot with 0.5 point highlighted.

## 7.6 Flap cleavage assay

The Okazakisome was reconstructed on DNA featuring the DNA structure generated during Okazaki fragment maturation process. In order to understand whether or not the 15 nt 5'-flap was cleaved by *Sso*Fen1, after reconstitution, we performed a cleavage assay. As shown in Figure 7.13, when the labelled DNA was incubated with either *Sso*Fen1 or with both *Sso*PCNA and *Sso*Fen1, a 28 nt cleavage product was detected. This clearly shows that it is likely that the reconstituted protein complex occurred on flap-cleaved DNA.

## 7.7 The 'tool–belt' model: when a single PCNA ring can load simultaneously three client proteins at once

To interpret our 3D–EM reconstruction, we used the structures calculated for its isolated components, and for the three subcomplexes *Hsa*PCNA–*Hsa*Fen1 [Sakurai et al., 2005], *Pfu*PCNA–*Pfu*PolB•DNA [Mayanagi et al., 2011] and *Hsa*PCNA–*Hsa*Lig1•DNA [Mayanagi et al., 2009]. This fitting could lead to two different Okazakisome architectures: one with *Sso*PCNA layered between its client proteins, the other one with *Sso*PCNA carrying the three client proteins at the front face. Sliding clamps have unique front and back faces [Mailand et al., 2013]. Although a recent paper showed that ubiquitin interacts with the back face of *Sce*PCNA [Freudenthal et al., 2010], it is generally believed that client proteins interact with the front face of PCNA [Mailand et al., 2013]. Indeed, the principal contact point on PCNA for client proteins, the inter–domain connector loop (IDCL), is found towards the front face of the PCNA ring. Notably, the IDCL–interacting motif within the ligase, which is required for functional interaction with PCNA, is found within a short internal loop in the ligase [Pascal et al., 2006]. Spatial constraints suggest it is unlikely that that *Sso*Lig1 could dock with the IDCL of *Sso*PCNA yet be located behind the ring. Nevertheless, to discriminate between the two possible map interpretations, we performed an activity assay in which we tested the preference of

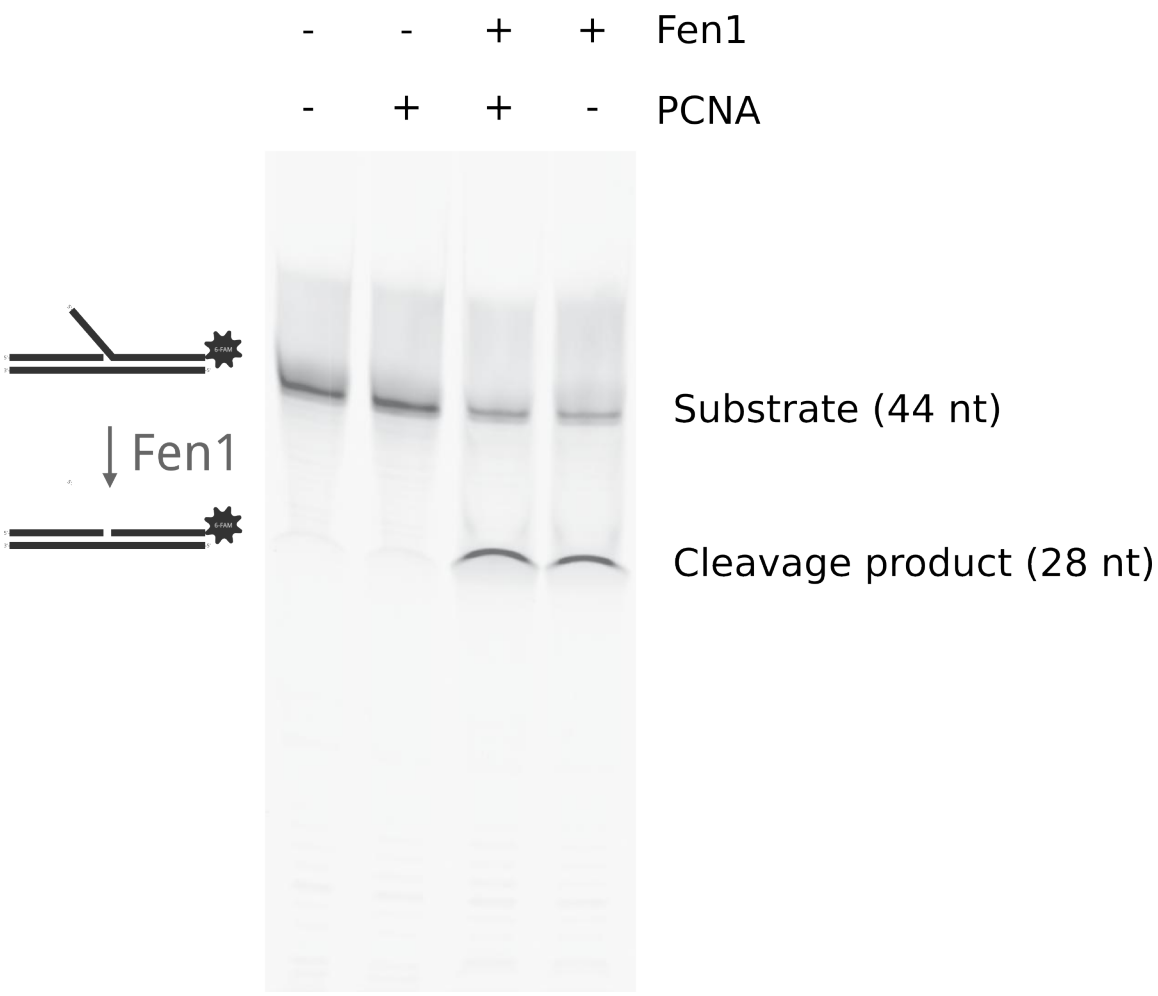



**Figure 7.13: *Sso*Fen1 flap cleavage assay.** Reactions contained the indicated proteins at 1.5 $\mu$M. Following termination of the cleavage reaction, samples were denatured and electrophoresed on a 12% denaturing polyacrylamide gel. The positions of migration of the substrate and cleavage product are indicated. The schematic to the left of the gel indicates the substrate and product with the 6-FAM (6-Carboxyfluorescein) fluorophore represented by a grey star (Experiment performed by Yuli Xu).

the ligase for either a front-face or a back-face ligation (Figure 7.17).

Additionally, we performed the GraFix procedure with DNA, *Sso*PolB1, *Sso*Fen1 and *Sso*PCNA in the absence of *Sso*Lig1. The resultant particles lack the density ascribed to *Sso*Lig1 in the full Okazakisome (Figure 7.16).

The fully assembled Okazakisome (Figure 7.14 and Figure 7.16) exhibits features compatible with previous EM studies of related sub-assemblies [Mayanagi et al., 2009, 2011]. To glean insight on the mechanism by which *Sso*PCNA coordinates three client proteins to process Okazaki fragments, we fitted each constituent of the assembly using Chimera, based on their volumes and shapes (Figure 7.14). It is not possible to determine unambiguously the trajectory of DNA within the complex. This could be because the DNA we designed to load one and only one copy of PCNA is completely embedded in the structure. To define the position of the sliding clamp within the map, we used crystallographic structures for the *Sso*PCNA ring (2NTI) [Hlinkova et al., 2008], for the *Hsa*PCNA–*Hsa*Fen1 complex (1UL1) [Sakurai et al., 2005] and for the *Sso*PCNA1–*Sso*PCNA2–*Sso*Fen1 (2IZO) [Doré et al., 2006]. 1UL1 and 2IZO facilitated docking and map interpretation since they have distinct asymmetric shapes. In 1UL1, each Fen1 molecule (chains X, Y and Z in the PDB file) crystallized in a distinct conformation. We tested fitting assemblies ABCX, ABCY and ABCZ from 1UL1. Based on the ligation experiment, we placed the PCNA ring at one end of the EM model, which exhibits a round shape. The complex is assembled on DNA, therefore the central cavity of PCNA is filled. The *Hsa*Fen1 Y chain from 1UL1 was the one that could better be accommodated in the EM density for *Sso*Fen1 (Figure 7.14). This is analogous to what was shown in the crystallographic study of the *Sso*PCNA1–*Sso*2 dimer associated to *Sso*Fen1, where the conformation of the endonuclease recalls the Y chain in 1UL1. We used the EM reconstructions for *Pfu*PCNA–*Pfu*LigI–DNA (EMDB-5220) and *Pfu*PCNA–*Pfu*PolB1•DNA (EMDB-1485) as probes to assign densities for PCNA-containing sub-complexes.

Combining the analysis of the map with the biochemical information available on the PCNA interaction with the client proteins [Beattie and Bell, 2012; Dionne et al., 2003, 2008], we modelled *Sso*PCNA2 interacting with the PIP box of *Sso*PolB1 and *Sso*PCNA3 with the PIP box in *Sso*Lig1. 1S5J fits extremely well in the density

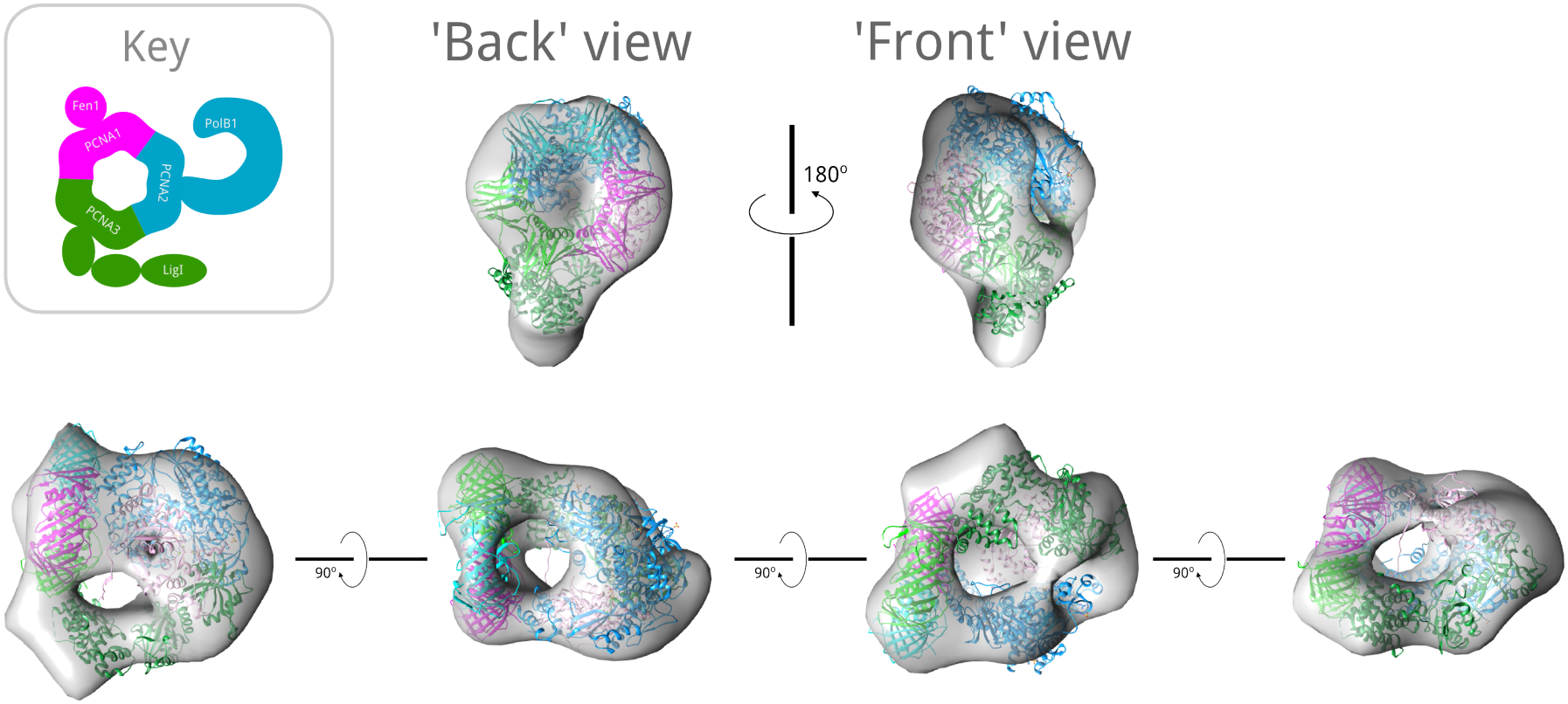



**Figure 7.14: Fitting the protein components within the *Sso*PCNA123–*Sso*PolB1–*Sso*Fen1–*Sso*LigI•DNA complex of *S. solfataricus*.** Top, 'front' and 'back' face of the Okazakisome 3D–EM model. Bottom, rotated sideviews. Individual protein components were fitted manually using Pymol.

connecting *Sso*PCNA2 and *Sso*PCNA3 (Figure 7.16 C), forming two extensive contacts with the *Sso*PCNA ring, consistent with the EM analysis of the *Pfu*PCNA–*Pfu*PolB complex loaded on DNA [Mayanagi et al., 2011]. We fitted the ligase at the front face of the *Sso*PCNA ring in our EM map using the crystallographic data for the full-length *S. solfataricus* protein (2HIV) [Pascal et al., 2006] and of hLigase233–919 in complex with DNA (1X9N) [Pascal et al., 2004]. A variety of studies have revealed that Ligase1 is a very flexible molecule [Pascal et al., 2006]. We performed its fitting placing the PIP–box containing–Nt close to *Sso*PCNA3, in the extended conformation of 2HIV, the crystal structure for the DNA-free ligase [Pascal et al., 2006]. The model shown in Figure 7.14 is in very good agreement with the 3D superposition of |*Pfu*PCNA123–*Pfu*PolI•DNA, *Sso*PCNA123–*Sso*Fen1 and *Pfu*PCNA123–*Pfu*Lig1•DNA (1UL1, EMDB-5220 and EMDB-1485). The main difference is the extended conformation of the ligase. The conformational malleability of Lig1 is highlighted by crystallographic structures which have revealed that ligase fully encircles the nicked DNA substrate to effect ligation. We speculate that when ligase adopts this final conformation it may displace DNA polymerase and flap-endonuclease, allowing their re–cycling to the next Okazaki fragment.

To further confirm the fitting of the ligase within the Okazakisome, we assembled a partial complex composed of *Sso*PCNA123–*Sso*PolB1–*Sso*Fen1 on DNA. We determined its three-dimensional structure using the EMD-5220 map as a starting model (Figure 7.15). We compared the *Pfu*PCNA–*Pfu*PolB•DNA starting model (Figure 7.16 B) with our *Sso*PCNA–*Sso*PolB1–*Sso*Fen1 assembly (Figure 7.16 B), and clearly saw a density (coloured pink in Figure 7.16) compatible with the speculated position of *Sso*Fen1 within the *Sso*PCNA ring. Fitting EMD-5220, 1UL1 (chain Y) and 2HIV in the Okazakisome (Figure 7.16 C) is compatible with *Sso*Lig1 in an extended conformation on the front face of the assembly.

## 7.8 Ligation assay

It is known that PCNA has a front-and a back-face [Mailand et al., 2013]. In our 3D–EM map of the Okazakisome, *Sso*PCNA123 could be fitted in the middle of the

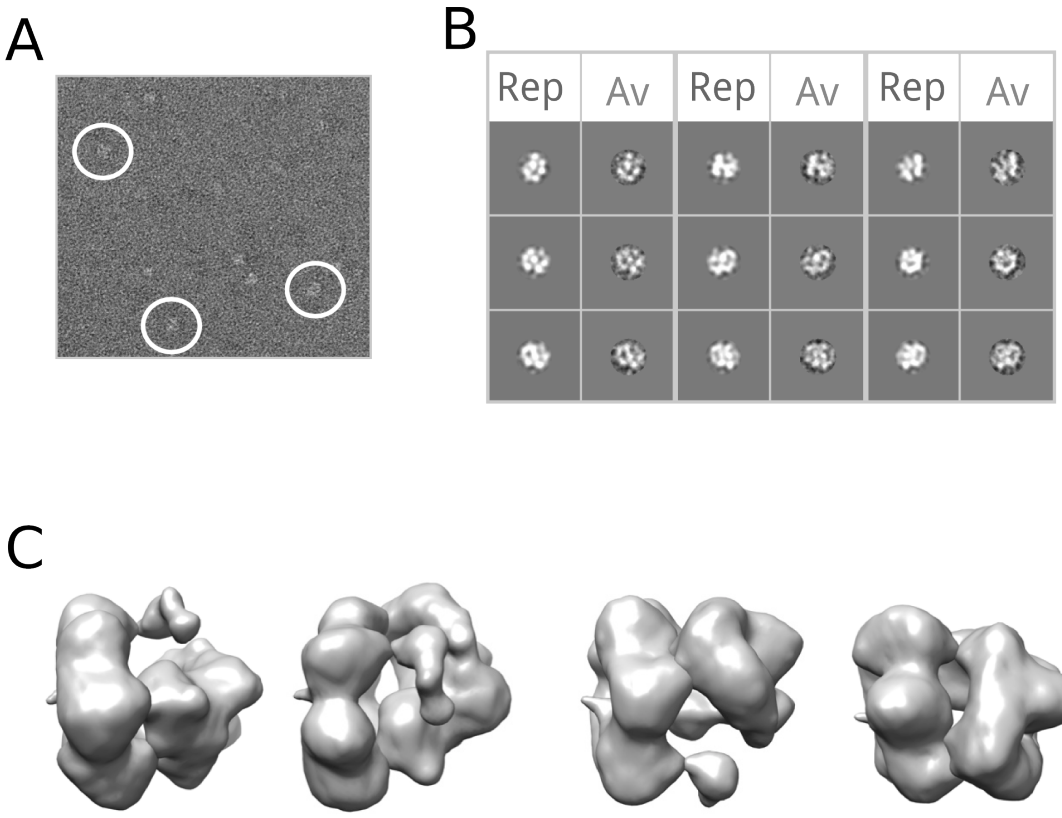



**Figure 7.15: Structural determination of the *Sso*PCNA123/*Sso*PolB1/*Sso*Fen1•DNA.** (A), Rappresentative micrograph. Single particles are shown in white circles. (B), Class averages (Av) and reprojections (Rep). (C), Surface representation of the *Sso*PCNA123/*Sso*PolB1/*Sso*Fen1•DNA assembly.

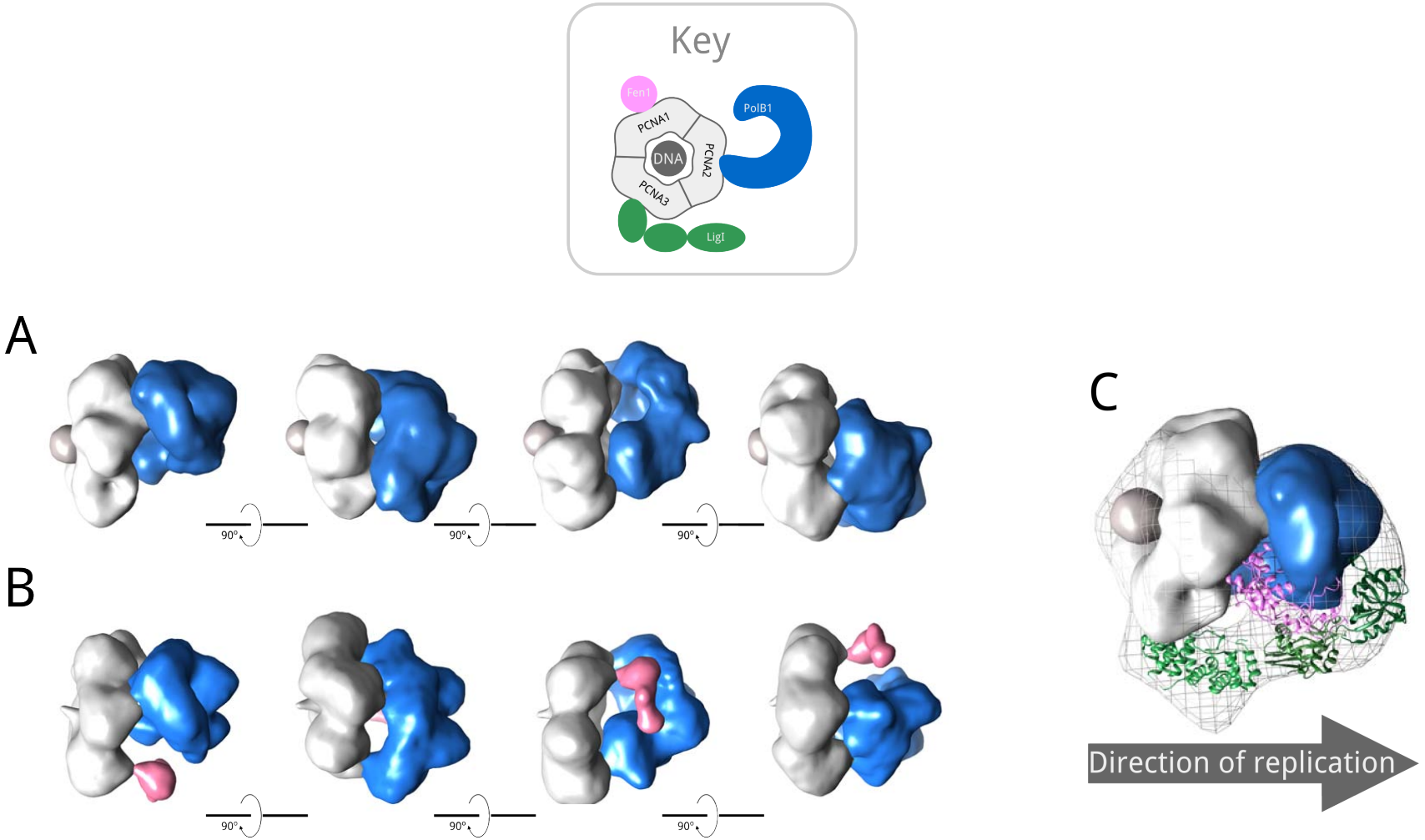



**Figure 7.16: Structural comparison of *Pfu*PCNA–*Pfu*PolB•DNA complex and *Sso*PCNA123–*Sso*PolB1–*Sso*Fen1•DNA complex of *S. solfataricus*.** (A) *Pfu*PCNA–*Pfu*PolB•DNA complex (EMD–5220). (B) Structural determination of *Sso*PCNA123–*Sso*PolB1–*Sso*PolB1•DNA complex. (C) Superimposition of *Pfu*PCNA–*Pfu*PolB•DNA complex (EMD–5220) with *Sso*PCNA123–*Sso*PolB1–*Sso*PolB1•DNA complex. *Sso*LigI is fitted in extended conformation within the Okazakisome.

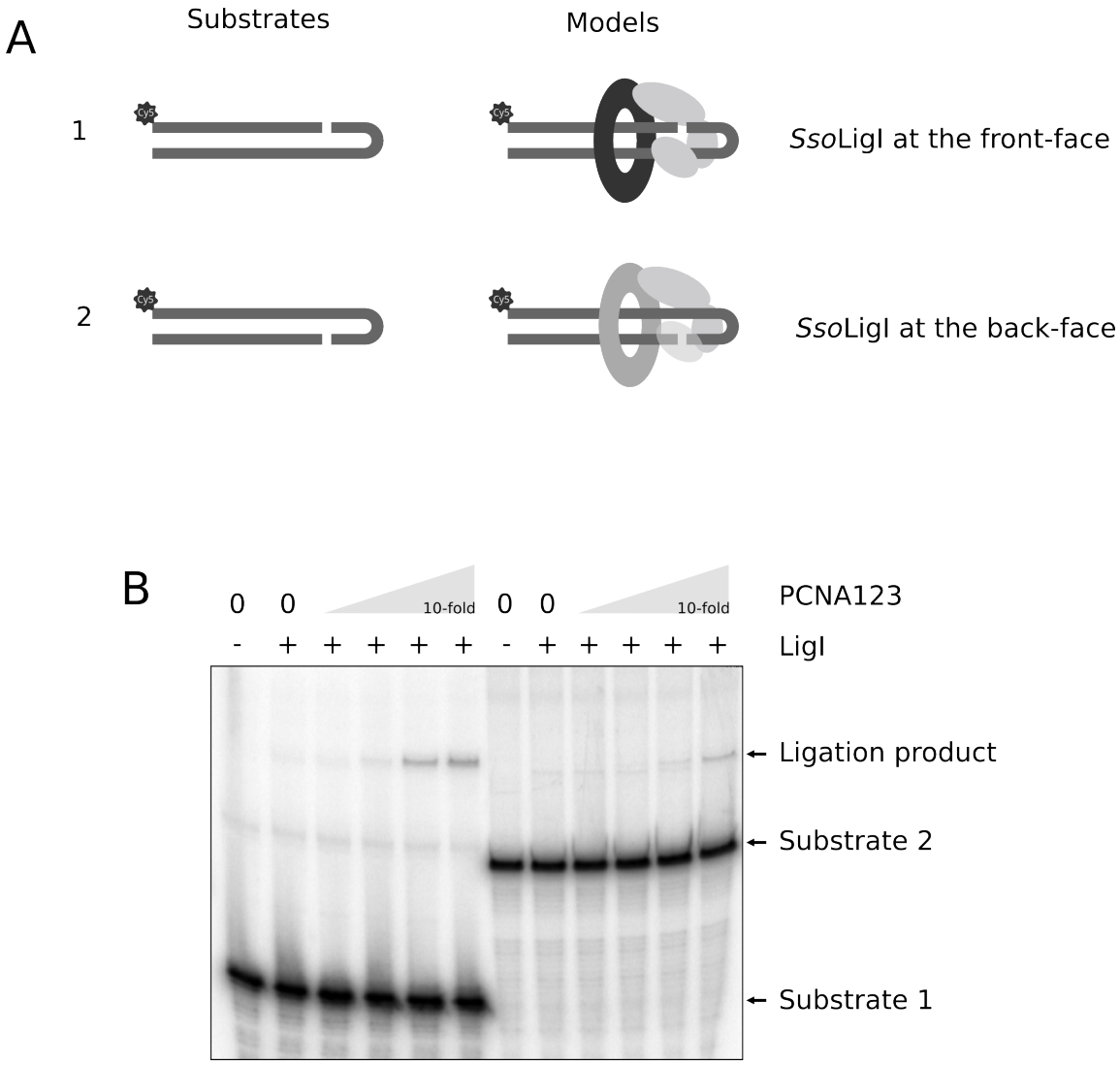



**Figure 7.17: *Sso*LigI ligation assay.** (A) Schematic of the substrates tested in order to biochemically determinate which face of *Sso*PCNA Schematics of the models (B) Ligation assay contained the indicated nicked hairpin DNA substrates, relative migration of substrates (1 and 2) and ligation products are indicated to the right of the gel image. Lanes 1 and 7 lacked protein. Lanes 2-5 and 8-11 contained 125 nM DNA *Sso*Lig1. Lanes 3 and 9, 4 and 10 and 5 and 11 had 12 nM, 1.2 nM and 0.12 nM *Sso*PCNA added (Experiment performed by Yuli Xu).

3D volume, placing *Sso*PolB1 and *Sso*Fen1 on the front-face of *Sso*PCNA123 while *Sso*Lig1 on the back–face. On the other hand, *Sso*PCNA could be fitted at the bottom of the 'bee–hive', placing the three clients proteins on the same face, the 'front–face' (Figure 7.14) . We designed two hairpin DNA substrates of the same overall sequence composition but with nicks on opposite strands. The positioning of the nick 9 base pairs from the hairpin limits the length of DNA available for interaction with *Sso*Lig1 and *Sso*PCNA. If *Sso*PCNA functionally interacts with *Sso*Lig1 via *Sso*PCNA's front-face then substrate 1 should reveal PCNA-stimulated ligation; if interaction with the back-face of *Sso*PCNA stimulates ligase then substrate 2 would show PCNA-enhancement of ligase activity. As seen in Figure 7.17, basal ligase activity on both substrates is equivalent, however, only substrate 1 shows PCNA-dependent stimulation of *Sso*Lig1 activity. Thus, our functional data support the fitting of all three client proteins on the front face of the *Sso*PCNA ring.

## 7.9 Concluding remarks

According to our structure, PCNA can load one client protein per subunit, forming a stable complex on DNA. The positioning of ligase in its extended conformation prevents steric clashes with the other client proteins. Our work reveals the answer to a long-debated biological question, showing that PCNA orchestrates the action of client proteins according to the 'toolbelt model', instead of loading just one client protein at one given time. A schematic model of Okazaki fragment maturation in *S. solfataricus* is presented in Figure 7.18.

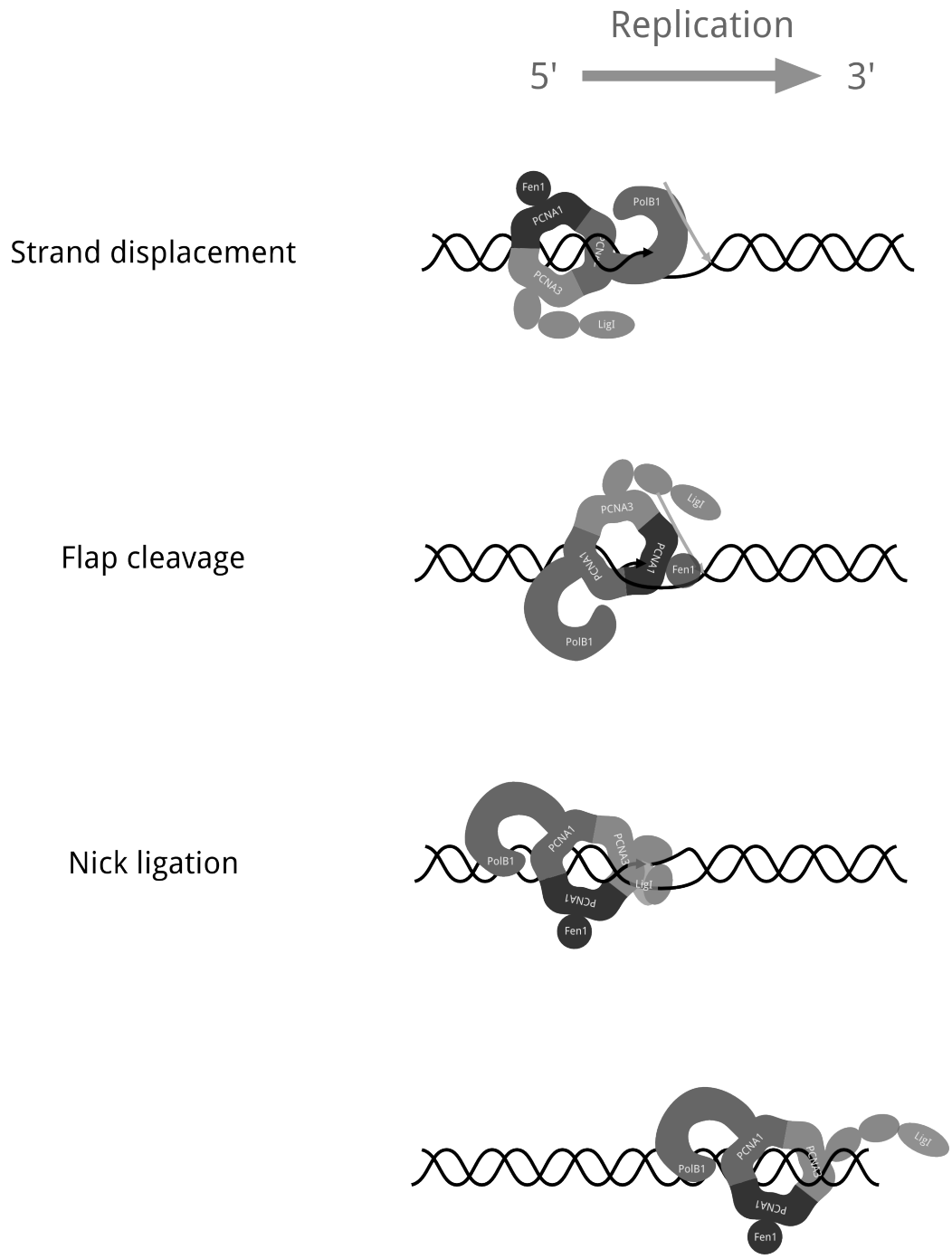



**Figure 7.18: 'Tools-belt' model of Okazaki fragment maturation in *S. solfataricus*.** Clients proteins *Sso*PolB1, *Sso*Fen1, and *Sso*LigI are shown to be interacting with their specific subunits composing the heterotrimeric *Sso*PCNA sliding clamp. During synthesis on the lagging strand, *Sso*PolB1 engages the template to perform DNA synthesis while *Sso*Fen1 and *Sso*Lig1 are 'carried' by *Sso*PCNA. As soon as the Okazakisome encounters a downstream Okazaki fragment, *Sso*PolB1 generates the 5'–flap by displacing the downstream Okazaki fragment (Strand displacement). *Sso*Fen1 binds to the specific flap structure and cleaves it, resulting in the formation of a nicked double stranded DNA (Flap cleavage). *Sso*LigI recognises the nick, encircles it and catalyses Okazaki fragment ligation (Nick ligation). The whole process, Okazaki fragment maturation, has imparted covalent stability to DNA double strand. Hence in this model, *Sso*PCNA orchestrate activities of *Sso*PolB1, *Sso*Fen1 and *Sso*LigI.

# Chapter 8

# Discussion and future perspectives

Genome sequencing and bioinformatic analysis of hyperthermophilic archaeon *P. abyssi* revealed an ORF (PAB2373) with sequence homology to previously reported MCM proteins [Cohen et al., 2003]. Since no biochemical and structural data are available for *Pab*MCM, the protein encoded by the PAB2373 (*Pab*MCM) was investigated in order to characterise its structure and function.

Comparison of the primary amino acid sequence of *Pab*MCM with three well–characterised archaeal MCM proteins revealed conserved residues in the N–terminal and the $AAA^+$ *Pab*MCM. At the N–terminal, A and B domain are less conserved if compared to the C domain. The C domain plays a key role in regulation of the helicase activity and in the communication between subunits within the ring [Kasiviswanathan et al., 2004; Barry et al., 2009; Sakakibara et al., 2008]. The $AAA^+$ catalytic domain is more conserved and contains structural motif typical of the active site of the $AAA^+$ proteins (Walker A and B, sensor 1 and 2, arginine finger). Additionally, the three $\beta$–hairpins, whose role in DNA binding and helicase activity has been determined by mutational analyses, are well conserved in *Pab*MCM [McGeoch et al., 2005; Jenkinson and Chong, 2003].

Bioinformatic analysis of the genomic sequence of *P. abyssi* reported the presence of two inteins, within the *Pab*mcm ORF. This is different from the *Pfu*mcm ORF, which contains only one intein. Inteins are found in all the three domains of life. Most intein host proteins seem to be present in the ancestors of eukaryotes and prokaryotes [Gogarten et al., 2002]. In archaeal genomes, inteins are found mostly in

gene encoding DNA replication proteins [Perler et al., 1997]. No evidence has been reported for an intein function that would be useful for their hosts. Additionally, gene replacement of an intein–containing homologue from a related organism showed no difference in the gene function [Frischkorn et al., 1998]. Protein sequences alignment showed that the first intein insertion, breaks apart C–terminal of the Walker A motif, whilst the second intein insertion falls into the glutamate switch, close to the $\beta$–hairpin H2I (Figure 4.2). The glutamate switch is a conserved residue whose role is to sense the presence of absence of a ligand (DNA in the case of MCM) Despite the sequence homology of *Pab*MCM with other archaeal MCM proteins, helicase assays with *Pab*MCM show low activity in the presence of a forked DNA substrate (G Henneke, pers. comm.). However, it has been showed that *Pab*MCM binds to dsDNA (Figure 4.16). In the near future, it would be interesting to investigate the ssDNA binding. Studies in *Pfu*MCM demonstrated that the helicase activity of MCM is stimulated by GINS [Yoshimochi et al., 2008]. However, in experiments where *Pab*GINS was added to the reaction with *Pab*MCM, no significant increment of the basal helicase activity was observed (G Henneke, pers. comm.). This indicates that other factors could be involved in activating the helicase activity of *Pab*MCM. Studies in *D. melanogaster* showed that MCM2-7 has no activity in the absence of GINS and Cdc45 [Ilves et al., 2010]. The archaeal Cdc45 candidate is GAN [Li et al., 2011]. Hence, similar mechanisms of activation could be involved in archaea. However, in previous work in *P. abyssi*, no interaction was observed between *Pab*GINS and *Pab*MCM. Surprisingly, recent proteomic studies revealed an unexpected association of *Pab*MCM with the archeal homologue of Xeroderma pigmentosum (XPD) [Pluchon et al., 2013]. *Pab*XPD has been cloned and a purification protocol is ongoing. Future work may reveal the molecular and structural basis of this interaction. It would be also interesting to investigate how GINS and GAN modulate the activity of MCM. Moreover, investigating the reason why *Sso*MCM and *Mth*MCM posses helicase activity in the presence of a forked DNA, whereas *Pab*MCM shows none, may reveal new mechanisms of regulation of the helicase activity. The helicase inactivity of *Pab*MCM may also be due to the buffer conditions. The unwinding activity of MCM2-7 is anion–dependent and it is inhibited by chlorides [Bochman

and Schwacha, 2008], while *Sso*MCM is not. Future investigation of the optimal buffer condition may provide an explanation of *Pab*MCM inactivity in isolation and in the current buffer.

EM studies showed that *Pab*MCM is a mixture of both single and double rings. SPA revealed that *Pab*MCM adopts at least three oligomeric states, with characteristic 2–, 7– and 8–fold symmetry (Figure 5.2, Figure 5.3). Previous EM studies of *Mth*MCM reported this behaviour and ascribed it to the salt concentration present in the preparation [Gómez-Llorente et al., 2005]. Recent biochemical evidence suggests that this behaviour is due to the effect of the temperature. In particular low temperature helps stabilising the double ring whereas higher temperature favours the single ring [Shin et al., 2009]. *P. abyssi* grows at temperature between 67°C 102°C under atmospheric pressure, with an optimum temperature of 96°C [Erauso et al., 1993]. These conditions cannot be used for protein purification due to the inability of the equipment to work at such high temperature. However, since a purification protocol in denaturing conditions was designed for purifying *Pab*MCM, it would be interesting to perform refolding at different temperatures and investigate if this may lead to an active helicase.

In this EM study a novel octameric single ring assembly is presented (Figure 5.5 A). Due to time constrains, only the docking of the crystal structure of the *Sso*MCM (Figure 1.4 B) and of the *Sso*MCM were performed (Figure 5.6). Rigid body fitting of the low resolution *Sso*MCM into the 3D–EM map suggests that two $\beta$–hairpins, PS1 and H2I, are pointing toward the central channel, while the third $\beta$–hairpin EXT fits with the density protruding between the subunits composing the ring (Figure 5.6). The same is also observed when rigid body fitting of the high resolution *Mth*MCM into the 3D–EM map, with the exception of the Ext hairpin. *Mth*MCM does not have the Ext hairpin. In future, other crystal structures of well–characterised $AAA^+$ proteins will be docked and their fitting investigated. Additional future analyses like pair–tilt validation will be carried out to assess the validity and the handness of this structure.

During the past few years, extensive biochemical analyses shed light on key roles of the three $\beta$–hairpins and NCL loop within the archaeal MCM complex. Docking

the crystal structure of the monomeric MCM protein into the available EM maps has led to model the archaeal MCM complex. However, no high resolution data of the archaeal MCM complex are available. Hence, high resolution structures of MCM helicase bound to DNA and in the presence and absence of ATP, ADP and hydrolysis transition state analogs such ADP–$Al_4$ are the next goal for understanding the structural basis of MCM helicase activity. In this respect, a dataset of ∼300 cryo–EM micrographs of *Pab*MCM and *Pab*MCM bound to 70bp blunt–end dsDNA is readly available for data processing and 3D–EM reconstruction.

Due to the antiparallel configuration of the dsDNA, the two strands of DNA are synthesized with different mechanisms. Replication occurs continuously on the leading strand, whilst the lagging strand is synthesised discontinuously, by short RNA–primed DNA fragments, the Okazaki fragments. To impart continuity to the lagging strand, RNA primers must be removed and replaced with DNA. DNA fragments must then be ligated together to ensure strand continuity.

The crenarchaeon *Sulfolobus solfataricus* possesses a simplified tool set for DNA replication compared to eukaryotes, and is therefore used as a model system for the study of replication proteins [Dionne et al., 2003]. *S. solfataricus* has a subset of the eukaryotic Okazaki fragment maturation factors, among which there are a heterotrimeric DNA sliding clamp, the proliferating cell nuclear antigen (PCNA), as well as DNA polymerase B1 (PolB1), the flap endonuclease, (Fen1), and an ATP–dependent DNA ligase (LigI). These proteins are necessary and sufficient for concerted DNA synthesis, RNA primer removal and strand ligation [Beattie and Bell, 2012]. In this biochemical scenario, PCNA plays a pivotal role. PCNA is a toroidal protein that encircles DNA and functions as a molecular platform at the replication fork to recruit numerous replication–associated enzymes. Biochemical and structural studies have highlighted a highly conserved mode of interaction between most PCNA–interacting proteins and PCNA. This interaction involves the recognition between a PCNA-interacting protein (PIP) domain on the client pro-

tein and the interdomain connector portion of each PCNA subunit, the IDCL loop. Most PCNA–binding proteins interact with its 'front' face, the side where replication would be occurring. However, a recent X-ray crystal structure of PCNA–Ubiquitin, in the context of translesion synthesis, showed that ubiquitin binds PCNA on its back face, opposite the IDCL loop, in an exceptional position for PCNA interactors [Freudenthal et al., 2010].

In *S. solfataricus*, PCNA is a heterotrimeric ring, where each subunit interacts preferentially with specific client proteins [Dionne et al., 2003]. As such, this is an ideal system to study specific PCNA–client protein interactions and complexes. In particular, in the context of Okazaki fragment maturation, *Sso*PCNA1 specifically interacts with *Sso*Fen1, *Sso*PCNA2 with the replicative *Sso*PolB1, and *Sso*PCNA3 with *Sso*Lig1 [Dionne et al., 2003]. *Sso* PCNA can also interact with a number of other proteins, among which the lesion–bypass polymerase Dpo4, a type 4 uracil DNA glycosylase and the DNA repair factor Xpf. It is still unclear how PCNA may coordinate the handoff of DNA from polymerase and Fen1 to DNA ligase and therefore coordinate the multistep process of Okazaki fragments maturation. Biochemical studies have put forward a 'molecular tool-belt' model to explain the mechanistic interaction of these proteins. In particular, a recent report by Beattie and Bell [2012] shows the PCNA–dependent *in vitro* reconstitution of DNA synthesis–dependent RNA primer removal and subsequent DNA ligation, providing supporting a model in which a single PCNA ring acts as the assembly platform for an Okazaki fragment maturation complex composed of PolB1, Fen1 and LigI. The heterotrimeric *Sso*PCNA provids us with a model to investigate the architectural basis of multi–enzyme coordination during lagging strand DNA replication.

We have assembled the PCNA–PolB1–Fen1–Lig complex on DNA and visualized it with electron microscopy coupled to single particle analysis. Our study highlights the mechanism that PCNA adopts to simultaneously load its client proteins, coordinating their function in time and space in a 'molecular tool belt' fashion. This is a long-standing model in the DNA replication community, but was never visualized so far. Indeed, we shown that the heterotrimeric *Sso*PCNA can load, simultaneously, three clients proteins per subunits forming a complex on DNA (Figure 7.14).

Although two possible fittings could be used for interpreting the 3D map, biochemical evidence (Figure 7.17) suggested that *Sso*LigI is on the front–face of the sliding clamp along with *Sso*PolB1, *Sso*LigI and *Sso*Fen1. The positioning of ligase in its extended conformation prevents steric clashes with the other client proteins (Figure 7.16).

The interaction PIP–boxes with IDCL domains, provide client proteins an extraordinary conformational flexibility between apparent 'carrier' and 'active' states which may provide a mechanism for ensuring accommodation of multiple different enzymes around a single PCNA, while preventing competitive access to DNA substrates [Beattie and Bell, 2012]. One example of such conformational switching is that of DNA ligase I, which in its active DNA–bound form, encircles DNA in a conformation which would presumably prevent access of other PCNA–bound factors to DNA [Pascal et al., 2004; Mayanagi et al., 2009]. In the absence of DNA, *Sso*LigI adopts an elongated form radiating from *Sso*PCNA [Pascal et al., 2006], thus providing an elegant explanation for how this particular protein may be 'carried' by PCNA without interfering with the action of *Sso*PolB1 and *Sso*Fen1 until an appropriate nick substrate arises.

Our work reveals the answer to a long–debated biological question, showing that PCNA orchestrates the action of client proteins according to the 'tool belt model', instead of loading just one client protein at one given time. Structural evidences revealed how a single PCNA ring can orchestrate multiple proteins. Sakurai et al. [2005] have directly visualised three molecules of human Fen1 bound to PCNA whereas Bubeck et al. [2011] reported the crystal structures of three RNase HII in complex with PCNA in *A. fulgidus*. However, no structural data have been reported for a single PCNA ring orchestrating multiple different proteins.

Finally, our studies of the heterotrimeric *Sso*PCNA in complex with *Sso*PolB1, *Sso*LigI and *Sso*Fen1 are likely to be informative regarding the mechanistic basis of action of the more complex eukaryotic homotrimeric PCNA and its client proteins (Figure 7.18). The orthologous factors to PolB1, Fen1 and Lig1 all interact with eukaryotic homotrimeric PCNA, and it is possible they may adopt the same arrangement.

# Bibliography

Agar, A. (1957). The measurement of the thickness of thin carbon films. *British Journal of Applied Physics*, 8(1):35.

Akita, M., Adachi, A., Takemura, K., Yamagami, T., Matsunaga, F., and Ishino, Y. (2010). Cdc6/Orc1 from *Pyrococcus furiosus* may act as the origin recognition protein and Mcm helicase recruiter. *Genes to cells*, 15(5):537–552.

Alois, J. and Rainer, H. (2009). Ion-Exchange Chromatography. *Methods in enzymology*, 463:349–371.

Andrew, P. (1970). Estimation of molecular size and molecular weights of biological compounds by gel filtration. *Methods of Biochemical Analysis, Volume 18*, pages 1–53.

Atanassova, N. and Grainge, I. (2008). Biochemical characterization of the minichromosome maintenance (MCM) protein of the crenarchaeote *Aeropyrum pernix* and its interactions with the origin recognition complex (ORC) proteins. *Biochemistry*, 47(50):13362–70.

Bae, B., Chen, Y.-H., Costa, A., Onesti, S., Brunzelle, J. S., Lin, Y., Cann, I. K. O., and Nair, S. K. (2009). Insights into the architecture of the replicative helicase from the structure of an archaeal MCM homolog. *Structure (London, England : 1993)*, 17(2):211–22.

Barry, E. and Bell, S. (2006). DNA replication in the archaea. *Microbiology and molecular biology reviews*, 70(4):876.

Barry, E., Lovett, J., Costa, A., Lea, S., and Bell, S. (2009). Intersubunit allosteric communication mediated by a conserved loop in the MCM helicase. *Proceedings of the National Academy of Sciences*, 106(4):1051.

Barry, E., McGeoch, A., Kelman, Z., and Bell, S. (2007). Archaeal MCM has separable processivity, substrate choice and helicase domains. *Nucleic acids research*, 35(3):988.

Beattie, T. R. and Bell, S. D. (2011). Molecular machines in archaeal DNA replication. *Current opinion in chemical biology*, 15(5):614–619.

Beattie, T. R. and Bell, S. D. (2012). Coordination of multiple enzyme activities by a single PCNA in archaeal Okazaki fragment maturation. *The EMBO journal*, 31(6):1556–1567.

Bell, S. D. and Botchan, M. R. (2013). The minichromosome maintenance replicative helicase. *Cold Spring Harbor perspectives in biology*, 5(11):a012807.

Bell, S. P. and Dutta, A. (2002). DNA replication in eukaryotic cells. *Annual review of biochemistry*, 71(1):333–374.

Blow, J. J. and Dutta, A. (2005). Preventing re-replication of chromosomal DNA. *Nature reviews. Molecular cell biology*, 6(6):476–86.

Bochman, M. L. and Schwacha, A. (2008). The MCM2-7 complex has in vitro helicase activity. *Molecular cell*, 31(2):287–93.

Bochman, M. L. and Schwacha, A. (2009). The MCM complex: unwinding the mechanism of a replicative helicase. *Microbiology and Molecular Biology Reviews*, 73(4):652–683.

Böttcher, B., Wynne, S., and Crowther, R. (1997). Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy. *Nature*.

Brewster, A., Wang, G., Yu, X., Greenleaf, W., Carazo, J., Tjajadi, M., Klein, M., and Chen, X. (2008). Crystal structure of a near-full-length archaeal MCM:

functional insights for an AAA+ hexameric helicase. *Proceedings of the National Academy of Sciences*, 105(51):20191.

Brochier-Armanet, C., Boussau, B., Gribaldo, S., and Forterre, P. (2008). Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nature Reviews Microbiology*, 6(3):245–252.

Bubeck, D., Reijns, M. A., Graham, S. C., Astell, K. R., Jones, E. Y., and Jackson, A. P. (2011). PCNA directs type 2 RNase H activity on DNA replication and repair substrates. *Nucleic acids research*, page gkq980.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., et al. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*, 273(5278):1058–1073.

Cann, I. K., Komori, K., Toh, H., Kanai, S., and Ishino, Y. (1998). A heterodimeric DNA polymerase: evidence that members of Euryarchaeota possess a distinct DNA polymerase. *Proceedings of the National Academy of Sciences*, 95(24):14250–14255.

Chen, H., Bjerknes, M., Kumar, R., and Jay, E. (1994). Determination of the optimal aligned spacing between the Shine–Dalgarno sequence and the translation initiation codon of *Escherichia coli* mRNAs. *Nucleic acids research*, 22(23):4953–4957.

Chen, Y., Yu, X., Kasiviswanathan, R., Shin, J., Kelman, Z., and Egelman, E. (2005). Structural polymorphism of *Methanothermobacter thermautotrophicus* MCM. *Journal of molecular biology*, 346(2):389–394.

Chen, Z., Speck, C., Wendel, P., Tang, C., Stillman, B., and Li, H. (2008). The architecture of the DNA replication origin recognition complex in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences*, 105(30):10326–10331.

Chevallet, M., Luche, S., and Rabilloud, T. (2006). Silver staining of proteins in polyacrylamide gels. *Nature protocols*, 1(4):1852–1858.

Chong, J. P., Hayashi, M. K., Simon, M. N., Xu, R.-M., and Stillman, B. (2000). A double-hexamer archaeal minichromosome maintenance protein is an ATP-dependent DNA helicase. *Proceedings of the National Academy of Sciences*, 97(4):1530–1535.

Cohen, G. N., Barbe, V., Flament, D., Galperin, M., Heilig, R., Lecompte, O., Poch, O., Prieur, D., Quérellou, J., Ripp, R., et al. (2003). An integrated analysis of the genome of the hyperthermophilic archaeon *Pyrococcus abyssi*. *Molecular microbiology*, 47(6):1495–1512.

Costa, A., Ilves, I., Tamberg, N., Petojevic, T., Nogales, E., Botchan, M. R., and Berger, J. M. (2011). The structural basis for MCM2–7 helicase activation by GINS and Cdc45. *Nature structural & molecular biology*, 18(4):471–477.

Costa, A., Pape, T., Van Heel, M., Brick, P., Patwardhan, A., and Onesti, S. (2006a). Structural basis of the Methanothermobacter thermautotrophicus MCM helicase activity. *Nucleic acids research*, 34(20):5829.

Costa, A., Pape, T., van Heel, M., Brick, P., Patwardhan, A., and Onesti, S. (2006b). Structural studies of the archaeal MCM complex in different functional states. *Journal of structural biology*, 156(1):210–219.

Costa, A., Van Duinen, G., Medagli, B., Chong, J., Sakakibara, N., Kelman, Z., Nair, S. K., Patwardhan, A., and Onesti, S. (2008). Cryo-electron microscopy reveals a novel DNA-binding site on the MCM helicase. *The EMBO journal*, 27(16):2250–2258.

De Falco, M., Ferrari, E., De Felice, M., Rossi, M., Hübscher, U., and Pisani, F. (2006). The human GINS complex binds to and specifically stimulates human DNA polymerase $\alpha$-primase. *EMBO reports*, 8(1):99–103.

De Felice, M., Esposito, L., Pucci, B., Carpentieri, F., De Falco, M., Rossi, M., and Pisani, F. (2003). Biochemical characterization of a CDC6-like protein from the crenarchaeon Sulfolobus solfataricus. *Journal of Biological Chemistry*, 278(47):46424.

DeLano, W. L. (2002). The PyMOL molecular graphics system.

Dionne, I., Brown, N. J., Woodgate, R., and Bell, S. D. (2008). On the mechanism of loading the PCNA sliding clamp by RFC. *Molecular microbiology*, 68(1):216–222.

Dionne, I., Nookala, R. K., Jackson, S. P., Doherty, A. J., and Bell, S. D. (2003). A Heterotrimeric PCNA in the Hyperthermophilic Archaeon *Sulfolobus solfataricus*. *Molecular cell*, 11(1):275–282.

Doré, A. S., Kilkenny, M. L., Jones, S. A., Oliver, A. W., Roe, S. M., Bell, S. D., and Pearl, L. H. (2006). Structure of an archaeal PCNA1-PCNA2-FEN1 complex: elucidating PCNA subunit and client enzyme specificity. *Nucleic acids research*, 34(16):4515–4526.

Dube, P., Tavares, P., Lurz, R., and Van Heel, M. (1993). The portal protein of bacteriophage SPP1: a DNA pump with 13-fold symmetry. *The EMBO journal*, 12(4):1303.

Dueber, E. L. C., Corn, J. E., Bell, S. D., and Berger, J. M. (2007). Replication origin recognition and deformation by a heterodimeric archaeal Orc1 complex. *Science*, 317(5842):1210–1213.

Duggin, I. G. and Bell, S. D. (2006). The chromosome replication machinery of the archaeon *Sulfolobus solfataricus*. *Journal of Biological Chemistry*, 281(22):15029–15032.

Edgell, D. and Doolittle, W. (1997). Archaea and the origin(s) of DNA replication proteins. *Cell*, 89(7):995–998.

Edler, K. (2013). Small-Angle Neutron Scattering. Lecture presentation at the $13^{th}$ Oxford School on Neutron Scattering.

Erauso, G., Reysenbach, A.-L., Godfroy, A., Meunier, J.-R., Crump, B., Partensky, F., Baross, J. A., Marteinsson, V., Barbier, G., Pace, N. R., et al. (1993). *Pyrococcus abyssi* sp. nov., a new hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent. *Archives of Microbiology*, 160(5):338–349.

Erzberger, J. P., Pirruccello, M. M., and Berger, J. M. (2002). The structure of bacterial DnaA: implications for general mechanisms underlying DNA replication initiation. *The EMBO journal*, 21(18):4763–73.

Feigin, L., Svergun, D. I., and Taylor, G. W. (1987). *Structure analysis by small-angle X-ray and neutron scattering*. Springer.

Fernández-Cid, A., Riera, A., Tognetti, S., Herrera, M. C., Samel, S., Evrin, C., Winkler, C., Gardenal, E., Uhle, S., and Speck, C. (2013). An ORC/Cdc6/MCM2-7 complex is formed in a multistep reaction to serve as a platform for MCM double-hexamer assembly. *Molecular cell*, 50(4):577–88.

Fletcher, R., Bishop, B., Leon, R., Sclafani, R., Ogata, C., and Chen, X. (2003). The structure and function of MCM from archaeal *M. thermoautotrophicum*. *Nature Structural & Molecular Biology*, 10(3):160–167.

Forsburg, S. L. (2004). Eukaryotic MCM proteins: beyond replication initiation. *Microbiology and Molecular Biology Reviews*, 68(1):109–131.

Frank, J. (1996). *Three-dimensional electron microscopy of macromolecular assemblies*. Academic Press.

Frank, J., Shimkin, B., and Dowse, H. (1981). SpiderâĂŤa modular software system for electron image processing. *Ultramicroscopy*, 6(4):343–357.

Freudenthal, B. D., Gakhar, L., Ramaswamy, S., and Washington, M. T. (2010). Structure of monoubiquitinated PCNA and implications for translesion synthesis and DNA polymerase exchange. *Nature structural & molecular biology*, 17(4):479–484.

Frigola, J., Remus, D., Mehanna, A., and Diffley, J. F. X. (2013). ATPase-dependent quality control of DNA replication origin licensing. *Nature*, 495(7441):339–43.

Frischkorn, K., Sander, P., Scholz, M., Teschner, K., Prammananan, T., and Böttger, E. C. (1998). Investigation of mycobacterial recA function: protein introns in the RecA of pathogenic mycobacteria do not affect competency for homologous recombination. *Molecular microbiology*, 29(5):1203–1214.

Gaberc-Porekar, V. and Menart, V. (2001). Perspectives of immobilized-metal affinity chromatography. *Journal of biochemical and biophysical methods*, 49(1):335–360.

Galagan, J. E., Nusbaum, C., Roy, A., Endrizzi, M. G., Macdonald, P., FitzHugh, W., Calvo, S., Engels, R., Smirnov, S., Atnoor, D., et al. (2002). The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity. *Genome research*, 12(4):532–542.

Gambus, A., Jones, R., Sanchez-Diaz, A., Kanemaki, M., Van Deursen, F., Edmondson, R., and Labib, K. (2006). GINS maintains association of Cdc45 with MCM in replisome progression complexes at eukaryotic DNA replication forks. *Nature cell biology*, 8(4):358–366.

Gambus, A., Van Deursen, F., Polychronopoulos, D., Foltman, M., Jones, R. C., Edmondson, R. D., Calzada, A., and Labib, K. (2009). A key role for Ctf4 in coupling the MCM2-7 helicase to DNA polymerase $\alpha$ within the eukaryotic replisome. *The EMBO journal*, 28(19):2992–3004.

Gaudier, M., Schuwirth, B. S., Westcott, S. L., and Wigley, D. B. (2007). Structural basis of DNA replication origin recognition by an ORC protein. *Science*, 317(5842):1213–1216.

Gilbert, D. M. (2001). Making sense of eukaryotic DNA replication origins. *Science*, 294(5540):96–100.

Glatter, O. (1977). A new method for the evaluation of small-angle scattering data. *Journal of Applied Crystallography*, 10(5):415–421.

Gogarten, J. P., Senejani, A. G., Zhaxybayeva, O., Olendzenski, L., and Hilario, E. (2002). Inteins: structure, function, and evolution. *Annual Reviews in Microbiology*, 56(1):263–287.

Gómez-Llorente, Y., Fletcher, R., Chen, X., Carazo, J., and Martín, C. (2005). Polymorphism and double hexamer structure in the archaeal minichromosome main-

tenance (MCM) helicase from *Methanobacterium thermoautotrophicum*. *Journal of Biological Chemistry*, 280(49):40909.

Grabowski, B. and Kelman, Z. (2003). Archaeal DNA replication: eukaryal proteins in a bacterial context. *Annual Reviews in Microbiology*, 57(1):487–516.

Graham, B. W., Schauer, G. D., Leuba, S. H., and Trakselis, M. A. (2011). Steric exclusion and wrapping of the excluded DNA strand occurs along discrete external binding paths during MCM helicase unwinding. *Nucleic acids research*, 39(15):6585–6595.

Grainge, I., Gaudier, M., Schuwirth, B. S., Westcott, S. L., Sandall, J., Atanassova, N., and Wigley, D. B. (2006). Biochemical Analysis of a DNA Replication Origin in the Archaeon *Aeropyrum pernix*. *Journal of molecular biology*, 363(2):355–369.

Guinier, A., Fournet, G., Walker, C. B., and Yudowitch, K. L. (1955). *Small-angle scattering of X-rays*. Wiley New York.

Gulbis, J. M., Kelman, Z., Hurwitz, J., O'Donnell, M., and Kuriyan, J. (1996). Structure of the C-Terminal Region of p21$^{WAF1/CIP1}$ Complexed with Human PCNA. *Cell*, 87(2):297–306.

Henneke, G., Flament, D., Hübscher, U., Querellou, J., and Raffin, J.-P. (2005). The hyperthermophilic euryarchaeota *Pyrococcus abyssi* likely requires the two DNA polymerases D and B for DNA replication. *Journal of molecular biology*, 350(1):53–64.

Herzog, F., Primorac, I., Dube, P., Lenart, P., Sander, B., Mechtler, K., Stark, H., and Peters, J.-M. (2009). Structure of the anaphase-promoting complex/cyclosome interacting with a mitotic checkpoint complex. *Science*, 323(5920):1477–1481.

Hlinkova, V., Xing, G., Bauer, J., Shin, Y. J., Dionne, I., Rajashankar, K. R., Bell, S. D., and Ling, H. (2008). Structures of monomeric, dimeric and trimeric PCNA: PCNA-ring assembly and opening. *Acta Crystallographica Section D: Biological Crystallography*, 64(9):941–949.

Hosfield, D. J., Frank, G., Weng, Y., Tainer, J. a., and Shen, B. (1998). Newly Discovered Archaebacterial Flap Endonucleases Show a Structure-Specific Mechanism for DNA Substrate Binding and Catalysis Resembling Human Flap Endonuclease-1. *Journal of Biological Chemistry*, 273(42):27154–27161.

Huber, H. and Prangishvili, D. (2006). Sulfolobales. In *The prokaryotes*, pages 23–51. Springer.

Hubscher, U., Maga, G., and Spadari, S. (2002). Eukaryotic DNA polymerases. *Annual review of biochemistry*, 71:133–63.

Hubscher, U. and Seo, Y.-S. (2001). Replication of the lagging strand: a concert of at least 23 polypeptides. *Molecules and cells*, 12(2):149–157.

Hwang, K. Y., Baek, K., Kim, H.-Y., and Cho, Y. (1998). The crystal structure of flap endonuclease-1 from *Methanococcus jannaschii. Nature structural biology*, 5(8):707–13.

Hyatt, M. (1981). Principles and techniques of electron microscopy. *Aspen, Rockville, MD*, page 73.

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular biology and evolution*, 2(1):13–34.

Ilves, I., Petojevic, T., Pesavento, J. J., and Botchan, M. R. (2010). Activation of the MCM2-7 helicase by association with Cdc45 and GINS proteins. *Molecular cell*, 37(2):247–258.

Ishimi, Y. (1997). A DNA helicase activity is associated with an MCM4,-6, and-7 protein complex. *Journal of Biological Chemistry*, 272(39):24508–24513.

Ishino, Y. and Ishino, S. (2012). Rapid progress of DNA replication studies in Archaea, the third domain of life. *Science China Life Sciences*, 55(5):386–403.

Jacob, F., Brenner, S., and Cuzin, F. (1963). On the regulation of DNA replication in bacteria. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 28, pages 329–348. Cold Spring Harbor Laboratory Press.

Jacques, D. A. and Trewhella, J. (2010). Small-angle scattering for structural biologyâĂŤExpanding the frontier while avoiding the pitfalls. *Protein Science*, 19(4):642–657.

Jenkinson, E. and Chong, J. (2003). Initiation of archaeal DNA replication. *Biochemical Society Transactions*, 31(Pt 3):669–673.

Jenkinson, E. R. and Chong, J. P. J. (2006). Minichromosome maintenance helicase activity is controlled by N- and C-terminal motifs and requires the ATPase domain helix-2 insert. *Proceedings of the National Academy of Sciences of the United States of America*, 103(20):7613–8.

Jeon, S.-J. and Ishikawa, K. (2003). A novel ADP-dependent DNA ligase from *Aeropyrum pernix* K1. *FEBS letters*, 550(1):69–73.

Jeruzalmi, D., O'Donnell, M., and Kuriyan, J. (2002). Clamp loaders and sliding clamps. *Current opinion in structural biology*, 12(2):217–224.

Johnson, A. and O'Donnell, M. (2005). Cellular DNA replicases: components and dynamics at the replication fork. *Annu. Rev. Biochem.*, 74:283–315.

Joyce, C. M. and Steitz, T. A. (1994). Function and structure relationships in DNA polymerases. *Annual review of biochemistry*, 63(1):777–822.

Kamada, K., Kubota, Y., Arata, T., Shindo, Y., and Hanaoka, F. (2007). Structure of the human GINS complex and its assembly and functional interface in replication initiation. *Nature Structural & Molecular Biology*, 14(5):388–396.

Kane, J. F. and Hartley, D. L. (1988). Formation of recombinant protein inclusion bodies in *Escherichia coli*. *Trends in biotechnology*, 6(5):95–101.

Kaplan, D. L., Davey, M. J., and O'Donnell, M. (2003). Mcm4, 6, 7 uses a âĂIJpump in ringâĂİ mechanism to unwind DNA by steric exclusion and actively translocate along a duplex. *Journal of Biological Chemistry*, 278(49):49171–49182.

Kasiviswanathan, R., Shin, J., Melamud, E., and Kelman, Z. (2004). Biochemical characterization of the *Methanothermobacter thermautotrophicus* minichromosome maintenance (MCM) helicase N-terminal domains. *Journal of Biological Chemistry*, 279(27):28358.

Kastner, B., Fischer, N., Golas, M. M., Sander, B., Dube, P., Boehringer, D., Hartmuth, K., Deckert, J., Hauer, F., Wolf, E., et al. (2007). GraFix: sample preparation for single-particle electron cryomicroscopy. *Nature methods*, 5(1):53–55.

Kearsey, S. E. and Labib, K. (1998). MCM proteins: evolution, properties, and role in DNA replication. *Biochimica et biophysica acta*, 1398(2):113–36.

Kelch, B. A., Makino, D. L., OâĂŹDonnell, M., and Kuriyan, J. (2011). How a DNA polymerase clamp loader opens a sliding clamp. *Science*, 334(6063):1675–1680.

Kelman, L. M. and Kelman, Z. (2003). MicroReview Archaea : an archetype for replication initiation studies? *Molecular microbiology*, 48:605–615.

Kelman, L. M. and Kelman, Z. (2004). Multiple origins of replication in archaea. *Trends in microbiology*, 12(9):399–401.

Kelman, Z., Lee, J., and Hurwitz, J. (1999). The single minichromosome maintenance protein of *Methanobacterium thermoautotrophicum* $\Delta$H contains DNA helicase activity. *Proceedings of the National Academy of Sciences*, 96(26):14783.

Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H., and Svergun, D. I. (2003). PRIMUS: a Windows PC-based system for small-angle scattering data analysis. *Journal of Applied Crystallography*, 36(5):1277–1282.

Kornberg, A. and Baker, T. A. (1992). *DNA replication*. WH Freeman San Francisco.

Kozak, M. (2005). Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, 361:13–37.

Krishna, T. S., Kong, X.-P., Gary, S., Burgers, P. M., and Kuriyan, J. (1994). Crystal structure of the eukaryotic DNA polymerase processivity factor PCNA. *Cell*, 79(7):1233–1243.

Krueger, S., Shin, J.-H., Curtis, J. E., Rubinson, K. a., and Kelman, Z. (2014). The solution structure of full-length dodecameric MCM by SANS and molecular modeling. *Proteins*, 82(10):1–11.

Kubota, Y., Takase, Y., Komori, Y., Hashimoto, Y., Arata, T., Kamimura, Y., Araki, H., and Takisawa, H. (2003). A novel ring-like complex of Xenopus proteins essential for the initiation of DNA replication. *Genes & development*, 17(9):1141.

Kühlbrandt, W. (2014). Biochemistry. The resolution revolution. *Science (New York, N.Y.)*, 343(6178):1443–4.

Labib, K. (2000). Uninterrupted MCM2-7 Function Required for DNA Replication Fork Progression. *Science*, 288(5471):1643–1647.

Labib, K., Kearsey, S., and Diffley, J. (2001). MCM2-7 proteins are essential components of prereplicative complexes that accumulate cooperatively in the nucleus during G1-phase and are required to establish, but not maintain, the S-phase checkpoint. *Molecular biology of the cell*, 12(11):3658.

Laemmli, U. K. et al. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *nature*, 227(5259):680–685.

Lee, J.-K. and Hurwitz, J. (2001). Processive DNA helicase activity of the minichromosome maintenance proteins 4, 6, and 7 complex requires forked DNA structures. *Proceedings of the National Academy of Sciences*, 98(1):54–59.

Leonard, A. C. and Méchali, M. (2013). DNA replication origins. *Cold Spring Harbor perspectives in biology*, 5(10):a010116.

Li, Z., Pan, M., Santangelo, T. J., Chemnitz, W., Yuan, W., Edwards, J. L., Hurwitz, J., Reeve, J. N., and Kelman, Z. (2011). A novel DNA nuclease is stimulated by association with the GINS complex. *Nucleic acids research*, 39(14):6114–23.

Lindås, A.-C. and Bernander, R. (2013). The cell cycle of archaea. *Nature Reviews Microbiology*, 11(9):627–638.

Lindert, S., Staritzbichler, R., Wötzel, N., KarakaÅ§, M., Stewart, P. L., and Meiler, J. (2009). EM-fold: De novo folding of alpha-helical proteins guided by intermediate-resolution electron microscopy density maps. *Structure (London, England : 1993)*, 17(7):990–1003.

Liu, C., Wu, R., Zhou, B., Wang, J., Wei, Z., Tye, B. K., Liang, C., and Zhu, G. (2012). Structural insights into the Cdt1-mediated MCM2–7 chromatin loading. *Nucleic acids research*, 40(7):3208–3217.

Liu, J., Smith, C. L., DeRyckere, D., DeAngelis, K., Martin, G. S., and Berger, J. M. (2000). Structure and function of Cdc6/Cdc18: implications for origin recognition and checkpoint control. *Molecular cell*, 6(3):637–648.

Liu, W., Pucci, B., Rossi, M., Pisani, F. M., and Ladenstein, R. (2008). Structural analysis of the *Sulfolobus solfataricus* MCM protein N-terminal domain. *Nucleic acids research*, 36(10):3235–3243.

Liu, Y., Kao, H.-I., and Bambara, R. A. (2004). Flap endonuclease 1: a central component of DNA metabolism. *Annual review of biochemistry*, 73(1):589–615.

Ludtke, S. J., Baldwin, P. R., and Chiu, W. (1999). EMAN: semiautomated software for high-resolution single-particle reconstructions. *Journal of structural biology*, 128(1):82–97.

Lundgren, M., Andersson, A., Chen, L., Nilsson, P., and Bernander, R. (2004). Three replication origins in Sulfolobus species: synchronous initiation of chromosome replication and asynchronous termination. *Proceedings of the National Academy of Sciences of the United States of America*, 101(18):7046.

Mailand, N., Gibbs-Seymour, I., and Bekker-Jensen, S. (2013). Regulation of PCNA–protein interactions for genome stability. *Nature reviews Molecular cell biology*, 14(5):269–282.

Maine, G., Sinha, P., and Tye, B. (1984). Mutants of *S. cerevisiae defective* in the maintenance of minichromosomes. *Genetics*, 106(3):365.

Maiorano, D., Lutzmann, M., and Méchali, M. (2006). MCM proteins and DNA replication. *Current opinion in cell biology*, 18(2):130–6.

Mann, M., Højrup, P., and Roepstorff, P. (1993). Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biological mass spectrometry*, 22(6):338–345.

Marinsek, N., Barry, E., Makarova, K., Dionne, I., Koonin, E., and Bell, S. (2006). GINS, a central nexus in the archaeal DNA replication fork. *EMBO reports*, 7(5):539–545.

Matsui, E. (1999). Thermostable Flap Endonuclease from the Archaeon, *Pyrococcus horikoshii*, Cleaves the Replication Fork-like Structure Endo/Exonucleolytically. *Journal of Biological Chemistry*, 274(26):18297–18309.

Matsui, E., Musti, K. V., Abe, J., Yamasaki, K., Matsui, I., and Harata, K. (2002). Molecular structure and novel DNA binding sites located in loops of flap endonuclease-1 from *Pyrococcus horikoshii*. *Journal of Biological Chemistry*, 277(40):37840–37847.

Matsunaga, F., Forterre, P., Ishino, Y., and Myllykallio, H. (2001). In vivo interactions of archaeal Cdc6/Orc1 and minichromosome maintenance proteins with the replication origin. *Proceedings of the National Academy of Sciences*, 98(20):11152–11157.

Matsunaga, F., Glatigny, A., Mucchielli-Giorgi, M.-H., Agier, N., Delacroix, H., Marisa, L., Durosay, P., Ishino, Y., Aggerbeck, L., and Forterre, P. (2007). Genomewide and biochemical analyses of DNA-binding activity of Cdc6/Orc1 and Mcm proteins in Pyrococcus sp. *Nucleic acids research*, 35(10):3214–3222.

Matsunaga, F., Norais, C., Forterre, P., and Myllykallio, H. (2003). Identification of short 'eukaryotic' Okazaki fragments synthesized from a prokaryotic replication origin. *EMBO reports*, 4(2):154–8.

Matsunaga, F., Takemura, K., Akita, M., Adachi, A., Yamagami, T., and Ishino, Y. (2010). Localized melting of duplex DNA by Cdc6/Orc1 at the DNA replication

origin in the hyperthermophilic archaeon *Pyrococcus furiosus*. *Extremophiles*, 14(1):21–31.

Matten, S. R., Schneider, T. D., Ringquist, S., and Brusilow, W. S. (1998). Identification of an intragenic ribosome binding site that affects expression of the uncB gene of the Escherichia coli proton-translocating ATPase (unc) operon. *Journal of bacteriology*, 180(15):3940–3945.

Mayanagi, K., Kiyonari, S., Nishida, H., Saito, M., Kohda, D., Ishino, Y., Shirai, T., and Morikawa, K. (2011). Architecture of the DNA polymerase B-proliferating cell nuclear antigen (PCNA)-DNA ternary complex. *Proceedings of the National Academy of Sciences*, 108(5):1845–1849.

Mayanagi, K., Kiyonari, S., Saito, M., Shirai, T., Ishino, Y., and Morikawa, K. (2009). Mechanism of replication machinery assembly as revealed by the DNA ligase-PCNA-DNA complex architecture. *Proceedings of the National Academy of Sciences*, 106(12):4647–4652.

McGeoch, A., Trakselis, M., Laskey, R., and Bell, S. (2005). Organization of the archaeal MCM complex on DNA and implications for the helicase mechanism. *Nature structural & molecular biology*, 12(9):756–762.

McNally, R., Bowman, G. D., Goedken, E. R., O'Donnell, M., and Kuriyan, J. (2010). Analysis of the role of PCNA-DNA contacts during clamp loading. *BMC structural biology*, 10(1):3.

Miller, M., Goto, R., Bernot, A., Zoorob, R., Auffray, C., Bumstead, N., and Briles, W. (2001). NEB, 1991–2001. pMAL Protein Fusion and Purification System Instruction Manual. *New England Biolads Inc., Version*, 5:1–33.

Moldovan, G.-L., Pfander, B., and Jentsch, S. (2007). PCNA, the maestro of the replication fork. *Cell*, 129(4):665–679.

Moreau, M. J., McGeoch, A. T., Lowe, A. R., Itzhaki, L. S., and Bell, S. D. (2007). ATPase site architecture and helicase mechanism of an archaeal MCM. *Molecular cell*, 28(2):304–314.

Morris, E. P., Rivera-Calzada, A., da Fonseca, P. C., Llorca, O., Pearl, L. H., and Spagnolo, L. (2011). Evidence for a remodelling of DNA-PK upon autophosphorylation from electron microscopy studies. *Nucleic acids research*, 39(13):5757–5767.

Mott, M. L. and Berger, J. M. (2007). DNA replication initiation: mechanisms and regulation in bacteria. *Nature Reviews Microbiology*, 5(5):343–354.

Moyer, S., Lewis, P., and Botchan, M. (2006). Isolation of the Cdc45/Mcm2-7/GINS (CMG) complex, a candidate for the eukaryotic DNA replication fork helicase. *Proceedings of the National Academy of Sciences*, 103(27):10236.

Myllykallio, H., Lopez, P., López-GarcìĄa, P., Heilig, R., Saurin, W., Zivanovic, Y., Philippe, H., and Forterre, P. (2000). Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. *Science*, 288(5474):2212–2215.

Nelson, D. L., Lehninger, A. L., and Cox, M. M. (2008). *Lehninger principles of biochemistry*. Macmillan.

Nishida, H., Kiyonari, S., Ishino, Y., and Morikawa, K. (2006). The Closed Structure of an Archaeal DNA Ligase from *Pyrococcus furiosus*. *Journal of molecular biology*, 360(5):956–967.

Nishida, H., Mayanagi, K., Kiyonari, S., Sato, Y., Oyama, T., Ishino, Y., and Morikawa, K. (2009). Structural determinant for switching between the polymerase and exonuclease modes in the PCNA-replicative DNA polymerase complex. *Proceedings of the National Academy of Sciences*, 106(49):20693–20698.

Ogura, T. and Wilkinson, A. (2001a). AAA+ superfamily ATPases: common structure–diverse function. *Genes to Cells*, 6(7):575–597.

Ogura, T. and Wilkinson, A. (2001b). AAA+ superfamily ATPases: common structure–diverse function. *Genes to Cells*, 6(7):575–597.

Olsen, G. J. and Woese, C. R. (1997). Archaeal genomics: an overview. *Cell*, 89(7):991–994.

Orlova, E. V. and Saibil, H. R. (2011). Structural analysis of macromolecular assemblies by electron microscopy. *Chemical reviews*, 111(12):7710–7748.

Oyama, T., Ishino, S., Fujino, S., Ogino, H., Shirai, T., Mayanagui, K., Saito, M., Nagasawa, N., Ishino, Y., and Morikawa, K. (2011). Architectures of archaeal GINS complexes, essential DNA replication initiation factors. *BMC biology*, 9(1):28.

Pacek, M., Tutter, A., Kubota, Y., Takisawa, H., and Walter, J. (2006). Localization of MCM2-7, Cdc45, and GINS to the site of DNA unwinding during eukaryotic DNA replication. *Molecular cell*, 21(4):581–587.

Pan, M., Kelman, L., and Kelman, Z. (2011). The archaeal PCNA proteins. *Biochemical Society Transactions*, 39(1):20.

Pape, T., Meka, H., Chen, S., Vicentini, G., Van Heel, M., and Onesti, S. (2003). Hexameric ring structure of the full-length archaeal MCM protein complex. *EMBO reports*, 4(11):1079–1083.

Pappin, D. J., Hojrup, P., and Bleasby, A. J. (1993). Rapid identification of proteins by peptide-mass fingerprinting. *Current biology*, 3(6):327–332.

Pascal, J. M., O'Brien, P. J., Tomkinson, A. E., and Ellenberger, T. (2004). Human DNA ligase I completely encircles and partially unwinds nicked DNA. *Nature*, 432(7016):473–478.

Pascal, J. M., Tsodikov, O. V., Hura, G. L., Song, W., Cotner, E. A., Classen, S., Tomkinson, A. E., Tainer, J. A., and Ellenberger, T. (2006). A flexible interface between DNA ligase and PCNA supports conformational switching and efficient ligation of DNA. *Molecular cell*, 24(2):279–291.

Perler, F. B., Olsen, G. J., and Adam, E. (1997). Compilation and analysis of intein sequences. *Nucleic acids research*, 25(6):1087–1093.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). UCSF Chimera visualization system for ex-

ploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–1612.

Pfeifer F, Palm P, S. K. (1994). *Molecular Biology of Archaea*. Gustav Fischer Verlag.

Pikuta, E. V., Hoover, R. B., and Tang, J. (2007). Microbial extremophiles at the limits of life. *Critical reviews in microbiology*, 33(3):183–209.

Plotkin, J. B. and Kudla, G. (2010). Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics*, 12(1):32–42.

Pluchon, P.-F., Fouqueau, T., Crezé, C., Laurent, S., Briffotaux, J., Hogrel, G., Palud, A., Henneke, G., Godfroy, A., Hausner, W., et al. (2013). An extended network of genomic maintenance in the archaeon *Pyrococcus abyssi* highlights unexpected associations between eucaryotic homologs. *PloS one*, 8(11):e79707.

Poplawski, A., Grabowski, B., Long, S. E., and Kelman, Z. (2001). The zinc finger domain of the archaeal minichromosome maintenance protein is required for helicase activity. *Journal of Biological Chemistry*, 276(52):49371–49377.

Randell, J. C. W., Bowers, J. L., Rodríguez, H. K., and Bell, S. P. (2006). Sequential ATP hydrolysis by Cdc6 and ORC directs loading of the Mcm2-7 helicase. *Molecular cell*, 21(1):29–39.

Remus, D., Beuron, F., Tolun, G., Griffith, J. D., Morris, E. P., and Diffley, J. F. X. (2009). Concerted loading of Mcm2-7 double hexamers around DNA during DNA replication origin licensing. *Cell*, 139(4):719–30.

Robertson, C. E., Harris, J. K., Spear, J. R., and Pace, N. R. (2005). Phylogenetic diversity and ecology of environmental Archaea. *Current opinion in microbiology*, 8(6):638–642.

Robinson, N., Dionne, I., Lundgren, M., Marsh, V., Bernander, R., and Bell, S. (2004). Identification of two origins of replication in the single chromosome of the archaeon *Sulfolobus solfataricus*. *Cell*, 116(1):25–38.

Robinson, N. P. and Bell, S. D. (2005). Origins of DNA replication in the three domains of life. *FEBS Journal*, 272(15):3757–3766.

Robinson, N. P. and Bell, S. D. (2007). Extrachromosomal element capture and the evolution of multiple replication origins in archaeal chromosomes. *Proceedings of the National Academy of Sciences*, 104(14):5806–5811.

Romero, A. and Garćia, P. (1991). Initiation of translation at AUC, AUA and AUU codons in Escherichia coli. *FEMS microbiology letters*, 84(3):325–330.

Rothenberg, E., Trakselis, M. a., Bell, S. D., and Ha, T. (2007). MCM forked substrate specificity involves dynamic interaction with the 5'-tail. *The Journal of biological chemistry*, 282(47):34229–34.

Sakakibara, N., Kasiviswanathan, R., Melamud, E., Han, M., Schwarz, F., and Kelman, Z. (2008). Coupling of DNA binding and helicase activity is mediated by a conserved loop in the MCM protein. *Nucleic Acids Research*, 36(4):1309.

Sakakibara, N., Kelman, L., and Kelman, Z. (2009a). How is the archaeal MCM helicase assembled at the origin? Possible mechanisms. *Biochemical Society Transactions*, 37(1):7–11.

Sakakibara, N., Kelman, L., and Kelman, Z. (2009b). Unwinding the structure and function of the archaeal MCM helicase. *Molecular microbiology*, 72(2):286–296.

Sakurai, S., Kitano, K., Yamaguchi, H., Hamada, K., Okada, K., Fukuda, K., Uchida, M., Ohtsuka, E., Morioka, H., and Hakoshima, T. (2005). Structural basis for recruitment of human flap endonuclease 1 to PCNA. *The EMBO journal*, 24:683–693.

Sambrook, J. and Russell, D. W. (2006). The Inoue Method for Preparation and Transformation of Competent *E. coli*:" Ultra Competent" Cells. *Cold Spring Harb Protoc*, 2:3944.

Sambrook, J., Russell, D. W., and Russell, D. W. (2001). *Molecular cloning: a laboratory manual (3-volume set)*, volume 999. Cold spring harbor laboratory press Cold Spring Harbor, New York:.

Samson, R. Y. and Bell, S. D. (2013). MCM loading–an open-and-shut case? *Molecular cell*, 50(4):457–8.

Sanchez-Pulido, L. and Ponting, C. (2011). Cdc45: The Missing RecJ Ortholog in Eukaryotes? *Bioinformatics*.

Sayers, J. R. and Artymiuk, P. J. (1998). Flexible loops and helical arches. *Nature Structural & Molecular Biology*, 5(8):668–670.

Schatz, M. and Van Heel, M. (1990). Invariant classification of molecular views in electron micrographs. *Ultramicroscopy*, 32(3):255–264.

Scheres, S. H. (2012). Relion: implementation of a bayesian approach to cryo-em structure determination. *Journal of structural biology*, 180(3):519–530.

Shin, J., Heo, G., and Kelman, Z. (2008). The *Methanothermobacter thermautotrophicus* Cdc6-2 protein, the putative helicase loader, dissociates the minichromosome maintenance helicase. *Journal of bacteriology*, 190(11):4091.

Shin, J.-H., Heo, G.-Y., and Kelman, Z. (2009). The *Methanothermobacter thermautotrophicus* MCM helicase is active as a hexameric ring. *The Journal of biological chemistry*, 284(1):540–6.

Shine, J. and Dalgarno, L. (1974). The 3âĂš-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proceedings of the National Academy of Sciences*, 71(4):1342–1346.

Shuman, S. (2009). DNA ligases: progress and prospects. *Journal of Biological Chemistry*, 284(26):17365–17369.

Simon, A. C., Zhou, J. C., Perera, R. L., van Deursen, F., Evrin, C., Ivanova, M. E., Kilkenny, M. L., Renault, L., Kjaer, S., Matak-Vinković, D., et al. (2014). A Ctf4 trimer couples the CMG helicase to DNA polymerase [agr] in the eukaryotic replisome. *Nature*.

Singleton, M. R., Morales, R., Grainge, I., Cook, N., Isupov, M. N., and Wigley, D. B. (2004). Conformational Changes Induced by Nucleotide Binding in

Cdc6/ORC From *Aeropyrum pernix*. *Journal of molecular biology*, 343(3):547–557.

Slaymaker, I. M., Fu, Y., Toso, D. B., Ranatunga, N., Brewster, A., Forsburg, S. L., Zhou, Z. H., and Chen, X. S. (2013). Mini-chromosome maintenance complexes form a filament to remodel DNA structure and topology. *Nucleic acids research*, 41(5):3446–56.

Slesarev, A. I., Mezhevaya, K. V., Makarova, K. S., Polushin, N. N., Shcherbinina, O. V., Shakhova, V. V., Belova, G. I., Aravind, L., Natale, D. A., Rogozin, I. B., et al. (2002). The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens. *Proceedings of the National Academy of Sciences*, 99(7):4644–4649.

Snider, J., Thibault, G., and Houry, W. A. (2008). The AAA+ superfamily of functionally diverse proteins. *Genome Biol*, 9(4):216.

Speck, C., Chen, Z., Li, H., and Stillman, B. (2005). ATPase-dependent cooperative binding of ORC and Cdc6 to origin DNA. *Nature structural & molecular biology*, 12(11):965–971.

Sriskanda, V., Kelman, Z., Hurwitz, J., and Shuman, S. (2000). Characterization of an ATP-dependent DNA ligase from the thermophilic archaeon *Methanobacterium thermoautotrophicum*. *Nucleic acids research*, 28(11):2221–2228.

Starmer, J. (2000). Free2Bind: tools for computing minimum free energy binding between two separate RNA molecules.

Stellwagen, E. (2009). Chapter 23 Gel Filtration. In Burgess, R. R. and Deutscher, M. P., editors, *Guide to Protein Purification, 2nd Edition*, volume 463 of *Methods in Enzymology*, pages 373–385. Academic Press.

Sun, J., Evrin, C., Samel, S. a., Fernández-Cid, A., Riera, A., Kawakami, H., Stillman, B., Speck, C., and Li, H. (2013). Cryo-EM structure of a helicase loading intermediate containing ORC-Cdc6-Cdt1-MCM2-7 bound to DNA. *Nature structural & molecular biology*, 20(8):944–51.

Sun, J., Kawakami, H., Zech, J., Speck, C., Stillman, B., and Li, H. (2012). Cdc6-induced conformational changes in ORC bound to origin DNA revealed by cryo-electron microscopy. *Structure*, 20(3):534–544.

Svergun, D. (1992). Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *Journal of Applied Crystallography*, 25(4):495–503.

Svergun, D. (1999). Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophysical journal*, 76(6):2879–2886.

Svergun, D., Barberato, C., and Koch, M. (1995). CRYSOL-a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *Journal of Applied Crystallography*, 28(6):768–773.

Szambowska, A., Tessmer, I., Kursula, P., Usskilat, C., Prus, P., Pospiech, H., and Grosse, F. (2014). DNA binding properties of human Cdc45 suggest a function as molecular wedge for DNA unwinding. *Nucleic acids research*, 42(4):2308–19.

Takahashi, K., Yamada, H., and Yanagida, M. (1994). Fission yeast minichromosome loss mutants mis cause lethal aneuploidy and replication abnormality. *Molecular biology of the cell*, 5(10):1145–58.

Takara, T. J. and Bell, S. P. (2011). Multiple Cdt1 molecules act at each origin to load replication-competent MCM2-7 helicases. *The EMBO journal*, 30(24):4885–4896.

Takayama, Y., Kamimura, Y., Okawa, M., Muramatsu, S., Sugino, A., and Araki, H. (2003). GINS, a novel multiprotein complex required for chromosomal DNA replication in budding yeast. *Genes & development*, 17(9):1153.

Takeda, D. and Dutta, A. (2005). DNA replication and progression through S phase. *Oncogene*, 24(17):2827–2843.

Tan, S. (2001). A Modular Polycistronic Expression System for Overexpressing Protein Complexes in *Escherichia coli*. *Protein expression and purification*, 21(1):224–234.

Tang, G., Peng, L., Baldwin, P. R., Mann, D. S., Jiang, W., Rees, I., and Ludtke, S. J. (2007). EMAN2: an extensible image processing suite for electron microscopy. *Journal of structural biology*, 157(1):38–46.

Tikole, S. and Sankararamakrishnan, R. (2006). A survey of mRNA sequences with a non-AUG start codon in RefSeq database. *Journal of Biomolecular Structure and Dynamics*, 24(1):33–41.

Tsutakawa, S. E., Classen, S., Chapados, B. R., Arvai, A. S., Finger, L. D., Guenther, G., Tomlinson, C. G., Thompson, P., Sarker, A. H., Shen, B., et al. (2011). Human flap endonuclease structures, DNA double-base flipping, and a unified understanding of the FEN1 superfamily. *Cell*, 145(2):198–211.

Turchi, J. J., Huang, L., Murante, R. S., Kim, Y., and Bambara, R. A. (1994). Enzymatic completion of mammalian lagging-strand DNA replication. *Proceedings of the National Academy of Sciences*, 91(21):9803–9807.

Urh, M., Simpson, D., and Zhao, K. (2009). Affinity chromatography: general methods. *Methods in enzymology*, 463:417–438.

Vallejo, L. F. and Rinas, U. (2004). Strategies for the recovery of active proteins through refolding of bacterial inclusion body proteins. *Microbial Cell Factories*, 3(1):11.

Van Heel, M. (1984). Multivariate statistical classification of noisy images (randomly oriented biological macromolecules). *Ultramicroscopy*, 13(1):165–183.

Van Heel, M. (1987). Angular reconstitution: a posteriori assignment of projection directions for 3d reconstruction. *Ultramicroscopy*, 21(2):111–123.

Van Heel, M. and Frank, J. (1981). Use of multivariate statistics in analysing the images of biological macromolecules. *Ultramicroscopy*, 6(2):187–194.

van Heel, M., Gowen, B., Matadeen, R., Orlova, E. V., Finn, R., Pape, T., Cohen, D., Stark, H., Schmidt, R., Schatz, M., et al. (2000). Single-particle electron cryo-microscopy: towards atomic resolution. *Quarterly reviews of biophysics*, 33(04):307–369.

van Heel, M., Harauz, G., Orlova, E. V., Schmidt, R., and Schatz, M. (1996). A new generation of the IMAGIC image processing system. *Journal of structural biology*, 116(1):17–24.

Volkov, V. V. and Svergun, D. I. (2003). Uniqueness of ab initio shape determination in small-angle scattering. *Journal of Applied Crystallography*, 36(3):860–864.

Waga, S. and Stillman, B. (1998). The DNA replicaiton fork in eukaryotic cells. *Annu. Rev. Biochem.*

Wang, H.-W., Noland, C., Siridechadilok, B., Taylor, D. W., Ma, E., Felderer, K., Doudna, J. A., and Nogales, E. (2009). Structural insights into rna processing by the human risc-loading complex. *Nature structural & molecular biology*, 16(11):1148–1153.

Warbrick, E. (2000). The puzzle of PCNA's many partners. *Bioessays*, 22(11):997–1006.

White, H. E., Saibil, H. R., Ignatiou, A., and Orlova, E. V. (2004). Recognition and separation of single particles with size variation by statistical analysis of their images. *Journal of molecular biology*, 336(2):453–460.

Wiedemann, C., Ohlenschlager, O., Medagli, B., Onesti, S., and Gorlach, M. (2014). NMR solution structure of the C-terminus of the minichromosome maintenance protein MCM from *Sulfolobus solfataricus*. To be published.

Williams, D. B. and Carter, C. B. (2009). The Transmission electron microscope. In *Transmission Electron Microscopy*, pages 3–22. Springer.

Williams, G. J., Johnson, K., Rudolf, J., McMahon, S. A., Carter, L., Oke, M., Liu, H., Taylor, G. L., White, M. F., and Naismith, J. H. (2006). Structure of the

heterotrimeric PCNA from *Sulfolobus solfataricus*. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, 62(10):944–948.

Woese, C. R. and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090.

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579.

Woodman, M. E. (2008). Direct PCR of intact bacteria (colony PCR). *Current protocols in microbiology*, pages A–3D.

Wriggers, W., Milligan, R. A., and McCammon, J. A. (1999). Situs: a package for docking crystal structures into low-resolution maps from electron microscopy. *Journal of structural biology*, 125(2):185–195.

Wu, Z., Liu, J., Yang, H., and Xiang, H. (2014). DNA replication origins in archaea. *Frontiers in microbiology*, 5(April):179.

Xiong, S., Zhang, L., and He, Q.-Y. (2008). Fractionation of Proteins by Heparin Chromatography. In Posch, A., editor, *2D PAGE: Sample Preparation and Fractionation*, volume 424 of *Methods in Molecular BiologyâĎć*, pages 213–221. Humana Press.

Yamagata, A., Kakuta, Y., Masui, R., and Fukuyama, K. (2002). The crystal structure of exonuclease RecJ bound to Mn2+ ion suggests how its characteristic motifs are involved in exonuclease activity. *Proceedings of the National Academy of Sciences*, 99(9):5908–5912.

Yanagi, K.-i., Mizuno, T., You, Z., and Hanaoka, F. (2002). Mouse geminin inhibits not only Cdt1-MCM6 interactions but also a novel intrinsic Cdt1 DNA binding activity. *Journal of Biological Chemistry*, 277(43):40871–40880.

Yao, N. Y. and O'Donnell, M. (2012). The RFC clamp loader: structure and function. In *The Eukaryotic Replisome: a Guide to Protein Structure and Function*, pages 259–279. Springer.

Yoshimochi, T., Fujikane, R., Kawanami, M., Matsunaga, F., and Ishino, Y. (2008). The GINS complex from *Pyrococcus furiosus* stimulates the MCM helicase activity. *Journal of Biological Chemistry*, 283(3):1601–1609.

Yu, X., VanLoock, M., Poplawski, A., Kelman, Z., Xiang, T., Tye, B., and Egelman, E. (2002). The *Methanobacterium thermoautotrophicum* MCM protein can form heptameric rings. *EMBO reports*, 3(8):792–797.

Zhang, X. and Wigley, D. B. (2008). The'glutamate switch'provides a link between ATPase activity and ligand binding in AAA+ proteins. *Nature structural & molecular biology*, 15(11):1223–1227.

Zillig, W., Stetter, K. O., Wunderl, S., Schulz, W., Priess, H., and Scholz, I. (1980). The Sulfolobus–'Caldariella' group: taxonomy on the basis of the structure of DNA-dependent RNA polymerases. *Archives of Microbiology*, 125(3):259–269.

Zou, L., Mitchell, J., Stillman, B., Zou, L. E. E., and Mitchell, J. A. Y. (1997). CDC45, a novel yeast gene that functions with the origin recognition complex and Mcm proteins in initiation of DNA replication. *Mol Cell Biol.*, 17(2).

# Appendices

## List of publications

Molecular Cell
Article

# Structure and Mechanism of the CMR Complex for CRISPR-Mediated Antiviral Immunity

Jing Zhang,[1,5] Christophe Rouillon,[1,5] Melina Kerou,[1] Judith Reeks,[1] Kim Brugger,[2] Shirley Graham,[1] Julia Reimann,[4] Giuseppe Cannone,[3] Huanting Liu,[1] Sonja-Verena Albers,[4] James H. Naismith,[1] Laura Spagnolo,[3,*] and Malcolm F. White[1,*]

[1]Biomedical Sciences Research Complex, University of St Andrews, Fife KY16 9ST, UK
[2]EASIH, University of Cambridge, Addenbrookes Hospital, Cambridge CB2 0QQ, UK
[3]Institute of Structural Molecular Biology and Centre for Science at Extreme Conditions, University of Edinburgh, Edinburgh EH9 3JR, UK
[4]Archaeal Molecular Biology Group, Max Planck Institute for Terrestrial Microbiology, Karl-von-Frisch-Strasse 10, 35043 Marburg, Germany
[5]These authors contributed equally to this work
*Correspondence: laura.spagnolo@ed.ac.uk (L.S.), mfw2@st-andrews.ac.uk (M.F.W.)
DOI 10.1016/j.molcel.2011.12.013

## SUMMARY

**The prokaryotic *c*lusters of *r*egularly *i*nterspaced *p*alindromic *r*epeats (CRISPR) system utilizes genomically encoded CRISPR RNA (crRNA), derived from invading viruses and incorporated into ribonucleoprotein complexes with CRISPR-associated (CAS) proteins, to target and degrade viral DNA or RNA on subsequent infection. RNA is targeted by the CMR complex. In *Sulfolobus solfataricus*, this complex is composed of seven CAS protein subunits (Cmr1-7) and carries a diverse "payload" of targeting crRNA. The crystal structure of Cmr7 and low-resolution structure of the complex are presented. *S. solfataricus* CMR cleaves RNA targets in an endonucleolytic reaction at UA dinucleotides. This activity is dependent on the 8 nt repeat-derived 5′ sequence in the crRNA, but not on the presence of a protospacer-associated motif (PAM) in the target. Both target and guide RNAs can be cleaved, although a single molecule of guide RNA can support the degradation of multiple targets.**

## INTRODUCTION

The CRISPR system has recently come to light as a complex mechanism of cell-mediated antiviral immunity (see Horvath and Barrangou, 2010; Karginov and Hannon, 2010; Marraffini and Sontheimer, 2010a for recent reviews). CRISPRs are genomically encoded arrays of short "spacer" sequences (20–72 bp, depending on the species), each flanked by a repeat sequence with an average length around 25–30 nt. CRISPR arrays are transcribed by the cellular RNA polymerase and processed to generate small crRNAs by nucleolytic cleavage within the repeat sequences (Brouns et al., 2008; Carte et al., 2008; Deltcheva et al., 2011). Processed crRNAs are utilized by a variety of CRISPR-associated (CAS) protein complexes as a guide RNA to target (and degrade the nucleic acid of) invading genetic entities with complementary nucleic acid sequences. This defensive "interference" process works in tandem with an adaptive "capture" process that allows the incorporation of new spacer sequences derived from viruses into the genomic CRISPR arrays.

The viral DNA sequences that become incorporated into CRISPR arrays are known as "protospacers" (Horvath et al., 2008). Protospacers are derived from both coding and noncoding regions of viral genomes, suggesting that the viral DNA, rather than RNA, is targeted by the process that captures new spacers (Horvath et al., 2008; Shah et al., 2009). Examination of the sequence context of protospacers revealed the presence of a conserved "protospacer-associated motif" (PAM) consisting of a di- or trinucleotide signature, immediately adjacent to the protospacer sequence (Bolotin et al., 2005; Deveau et al., 2008; Horvath et al., 2008; Mojica et al., 2009). The presence of a PAM is important for the recognition and restriction of invading mobile DNA elements (Deveau et al., 2008; Gudbergsdottir et al., 2011).

CAS protein complexes have recently been classified into three main subtypes (Makarova et al., 2011b). In *Escherichia coli*, the type I-E complex, commonly known as "CASCADE," consists of five protein subunits (CasA-E). CASCADE processes CRISPR transcripts into ~57 nt crRNAs and uses them to recognize invading viral DNA, which is subsequently cleaved by Cas3 (Brouns et al., 2008). A similar complex (type I-A) with several conserved subunits has been described in *Sulfolobus solfataricus* (Lintner et al., 2011). Many archaea and some bacteria also encode a type III-B system, known as the CMR complex (Haft et al., 2005). In the euryarchaeon *Pyrococcus furiosus*, the Cmr1-6 proteins have been purified as a complex that uses crRNA to target RNA (presumably viral mRNA in vivo), cleaving it with a molecular ruler mechanism guided by the 3′ end of the crRNA (Hale et al., 2009).

Here, we report the purification and characterization of the CMR complex from *S. solfataricus*. There are seven subunits, comprising Cmr1-7 and a crRNA component. Deep sequencing reveals a biased composition for the crRNA, which is largely derived from 2 of the 6 CRISPR loci. The crystal structure of the Cmr7 subunit has been solved and consists of a protein fold with a conserved surface that may mediate molecular

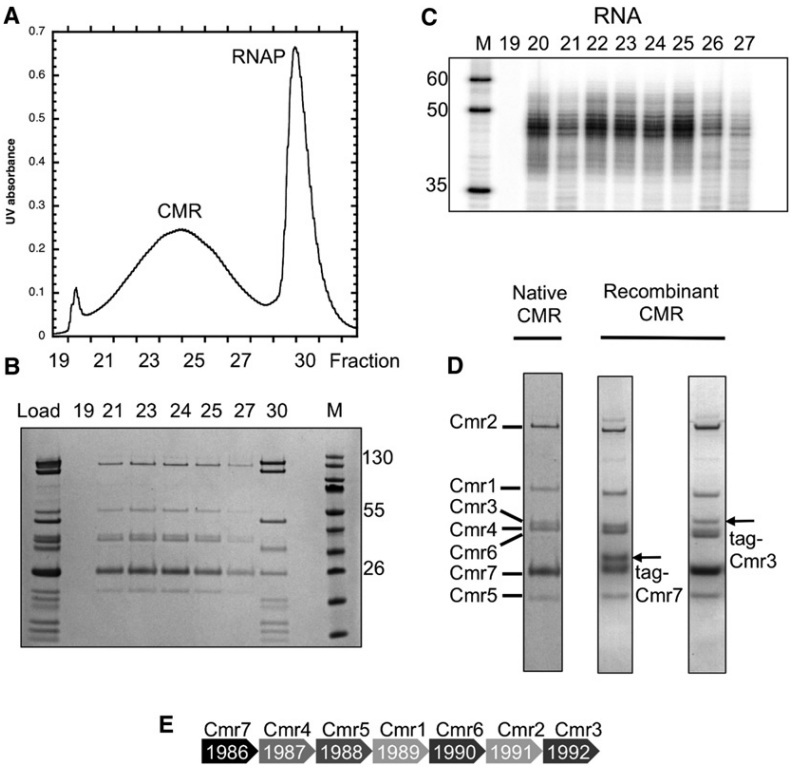



**Figure 1. Purification of the CMR Complex of *S. solfataricus***

(A) UV trace showing fractions eluting from final MonoQ column, with CMR and RNAP complexes resolved.

(B) SDS-PAGE analysis of fractions from MonoQ column, showing separation of RNAP (fraction 30) from the 7-subunit CMR complex.

(C) Denaturing gel electrophoresis of end-labeled nucleic acid reveals the presence of RNA copurifying with the CMR complex. The size range centered on 46 nt corresponds to a spacer with an 8 nt repeat-derived 5′ tag.

(D) Comparison of native and tagged versions of the CMR complex purified from *S. solfataricus*. Both tagged and untagged versions of Cmr7 are apparent, reflecting its higher stoichiometry in the complex.

(E) Mapping of Cmr1–7 onto the gene locus *sso1986* to *sso1992*.

interactions. The EM structure of the full CMR complex and a defined subcomplex are presented. We demonstrate that *S. solfataricus* CMR (SsoCMR) utilizes a sequence-dependent RNA cleavage mechanism without a molecular ruler.

## RESULTS

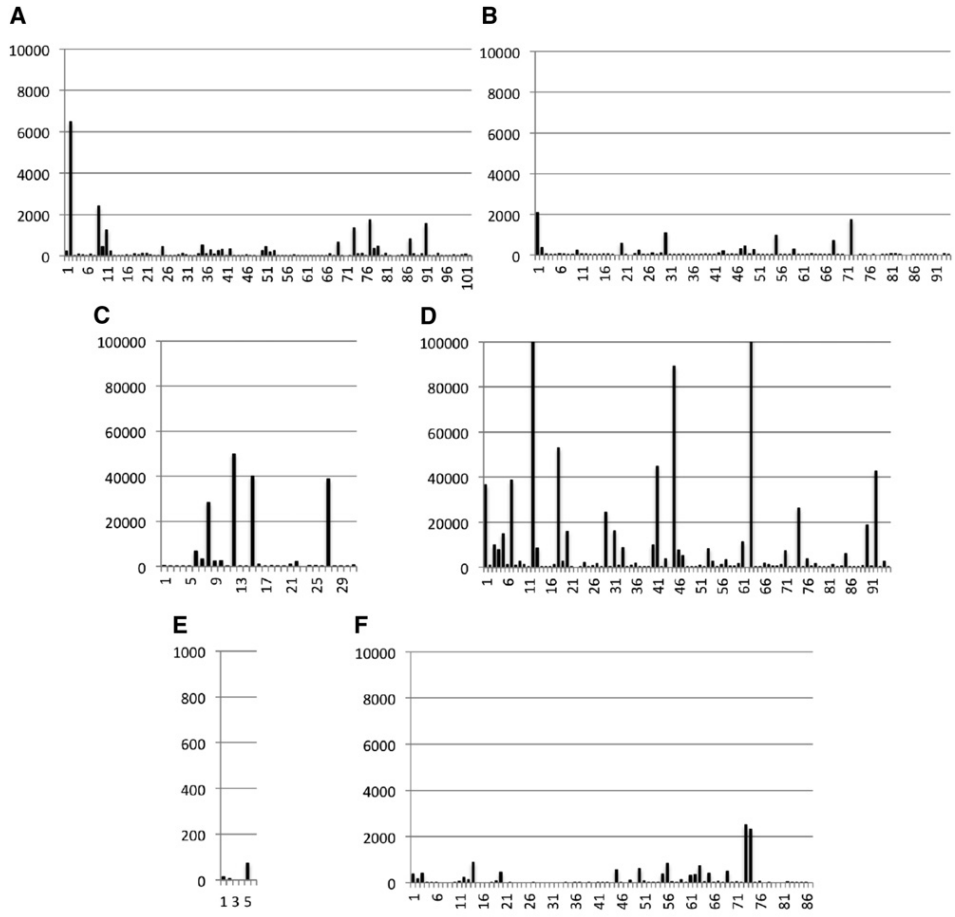
### Purification of the CMR Complex from *S. solfataricus*

The native CMR complex was purified from *S. solfataricus* using four sequential column chromatography steps (heparin, gel filtration, MonoS, and MonoQ). At each stage, column fractions were checked for the presence of the Cmr7 subunit by "dot blot" with a specific polyclonal antibody. SsoCMR copurified with the cellular RNA polymerase through the first three columns and was separated by the final anion exchange step (Figure 1A). The purified complex contained seven subunits corresponding to the products of genes *sso*1986 through *sso*1992. Subunits 1–6 were judged present at a 1:1 stoichiometry; densitometric analysis suggested that the Cmr7 subunit was present at a stoichiometry of three dimers in each complex (Figures 1 and S1). This was consistent with an overall size for the complex of 415 kDa (or 430 kDa including the RNA component), explaining the copurification on gel filtration with the 410 kDa RNA polymerase. The presence of RNA, with a variable fragment size centered on 46 nt, was confirmed (Figure 1C). This was in good agreement with the size of the crRNA species isolated from *P. furiosus* CMR (PfuCMR) (Hale et al., 2009), consistent with the presence of a spacer sequence of variable length with a CRISPR repeat-derived 8 nt tag at the 5′ end. By expressing subunits (Cmr7 or Cmr3) of SsoCMR with a polyhistidine tag in *S. solfataricus*, the intact complex could be isolated by incorporating an affinity chromatography step (Figure 1D).

### RNA Content of the CMR Complex

To determine the specific characteristics of the RNA component of SsoCMR, isolated RNA was cloned and deep-sequenced. We mapped 1.88 million sequence reads of 36 nt length onto the *S. solfataricus* spacers (Table S1). Analysis of the start and stop positions of the sequence reads for a subset of highly represented spacers from the A and D loci revealed the presence of sequence corresponding to the 5′ tag derived from the CRISPR repeat, with a clear demarcation at the eighth nucleotide, which corresponds to the site of cleavage by Cas6 (Carte et al., 2008; Lintner et al., 2011) (Figure S2). The 3′ ends of the sequenced RNA were more variable. Some spacers such as A2 and D43 displayed a short 3′ handle, while others appeared to have very little repeat-derived sequence at the 3′ end. Overall, this fits the suggestion that crRNA is processed by the Cas6 endoribonuclease followed by exonucleolytic digestion of the 3′ end (Hale et al., 2009). By contrast, crRNA isolated from the *S. solfataricus* CASCADE complex still includes the 3′ repeat-derived sequence (Lintner et al., 2011), suggesting that crRNAs are differentially processed depending on their ultimate destination in CASCADE or CMR.

*S. solfataricus* P2 has six CRISPR loci, denoted A–F, ranging from 8 to 103 repeats in length (Lillestøl et al., 2006). Whole-genome transcription data suggested that loci A, B, C, and D are all highly transcribed, while E and F are very weakly transcribed (Wurtzel et al., 2010). Examples of spacers derived from all six loci were detected in the CMR complex, but the distribution was highly biased. The majority of CMR-bound crRNAs were derived from locus D, followed by locus C, with significant underrepresentation from the other four loci (Figure 2). Within a locus, spacer representation in the CMR complex was highly variable, with numbers of sequenced crRNAs from adjacent spacers frequently differing by several orders of




| CRISPR locus | No. of spacers | Repeat | % of total spacers | % of crRNAs in CMR | Ratio observed/ expected |
|---|---|---|---|---|---|
| A | 102 | GATTAATCCCAAAAGGAATTGAAAG | 24 | 1.3 | 0.05 |
| B | 94 | GATTAATCCCAAAAGGAATTGAAAG | 23 | 0.6 | 0.03 |
| C | 31 | GATAATCTCTTATAGAATTGAAAG | 7.4 | 9.6 | 1.3 |
| D | 95 | GATAATCTCTTATAGAATTGAAAG | 23 | 88 | 3.8 |
| E | 7 | GATAATCTACTATAGAATTGAAAG | 1.7 | 0.005 | 0.003 |
| F | 88 | GCTAATCTACTATAGAATTGAAAG | 21 | 0.7 | 0.03 |

**Figure 2. Distribution of crRNA Bound by the *S. solfataricus* CMR Complex**

(A–F) Examples of crRNAs from all six CRISPR loci were observed, with a clear bias toward locus D, followed by locus C. The individual plots for each locus show that crRNA representation was highly variable, with adjacent spacers represented at levels that often varied by several orders of magnitude. For each graph, the x axis represents the position of each spacer in the locus and the y axis represents the number of sequenced matches to each spacer. 88% of the total sequence reads were derived from locus D, which represents a nearly 4-fold overrepresentation compared to the proportion of crRNAs encoded by that locus. CRISPR loci E and F, which are poorly transcribed, were significantly underrepresented, as expected. However, crRNAs from CRISPRs A and B, which are highly transcribed and thought to be actively adding spacers in vivo, are also significantly underrepresented in the CMR complex. The table shows the properties and representation in CMR of each CRISPR locus.

magnitude (Figure 2). In particular, one spacer (D63) accounted for 45% of the sequence reads, and the top ten spacers accounted for 79% of the total reads mapped to spacers (Table S2). Reverse transcripts of CRISPR arrays have been detected in *Sulfolobus* species (Lillestøl et al., 2009), but their significance remains unclear. Only 209 RNAs corresponding to reverse transcripts were sequenced, corresponding to 0.01% of the total (Table S1).

## The Crystal Structure of Cmr7

The crystal structure of Cmr7 (Sso1986) was solved to 2.05 Å resolution (PDB 2X5Q) (Figure 3). Sso1986 exhibits a fold consisting of six β sheets and four α helices, forming a dimer with a buried surface area of 1177 Å$^2$ and a concave face (Figures 3 and S3A). Homologs of Cmr7 are only detectable in the *Sulfolobales*, but given the rapid evolution of the CRISPR system, it is possible that distant homologs sharing this fold exist

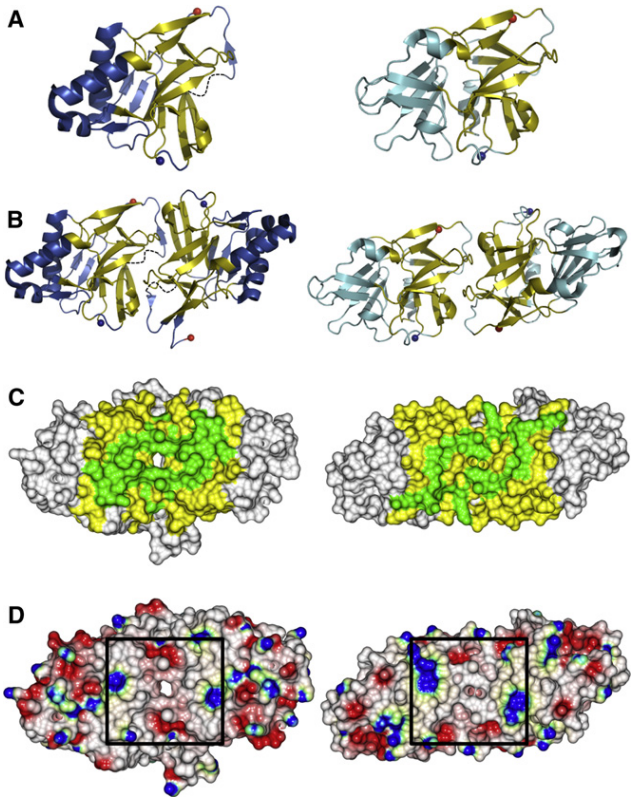



**Figure 3. Crystal Structures of Two Members of the Cmr7 Family, Viewed from the Concave Face of the Dimer**

(A) The Cmr7 proteins Sso1986 (left) and Sso1725 (right) contain a structurally conserved core (yellow) and a variable region (blue and cyan for Sso1986 and Sso1725, respectively). The β13-β14 loop of Sso1986 is disordered and is represented as a dashed, black line. The N and C termini are represented as blue and red spheres, respectively.

(B) Sso1986 and Sso1725 both form dimers, and the structurally conserved core is located at the dimer interface. The interface itself is also conserved between the two proteins.

(C) The structurally conserved residues (green) and secondary structure (yellow) are located close to the dimer interface with a significant proportion positioned at the concave face.

(D) Electrostatic surface images show that the regions of the concave face proximal to the dimer interface in both proteins (black box) have broad similarities.

**Table 1. Data Collection and Refinement Statistics for Sso1725**

| Data collection | Native | Anomalous |
|---|---|---|
| Wavelength (Å) | 0.97 | 1.60 |
| Space group | $P2_12_12_1$ | $P2_12_12_1$ |
| *a*, *b*, *c* (Å) | 77.75, 90.29, 111.65 | 77.71, 91.08, 111.94 |
| α, β, γ (°) | 90.0, 90.0, 90.0 | 90.0, 90.0, 90.0 |
| Resolution (Å) | 47.48–2.08 (2.13–2.08) | 77.11–2.54 (2.61–2.54) |
| I/σI | 19.2 (2.2) | 20.7 (2.8) |
| $R_{merge}$ | 0.04 (0.57) | 0.06 (0.71) |
| Completeness | 97.8 (86.1) | 99.9 (99.8) |
| Multiplicity | 4.7 (3.7) | 7.8 (7.9) |
| Anomalous completeness | - | 99.9 (99.7) |
| Anomalous multiplicity | - | 4.1 (4.1) |
| Refinement | Sso1725 | |
| $R_{work}/R_{free}$ | 0.199/0.231 | |
| Mean B value (Å$^2$) | | |
| All atoms | 59.442 | |
| Protein | 59.481 | |
| Water | 54.758 | |
| Root-mean-square deviations | | |
| Bond lengths (Å) | 0.01 | |
| Angles (°) | 1.31 | |

Each data set was collected on a single crystal at 100 K. Statistics are presented as averages with values for the highest-resolution shell in parentheses. $R_{free}$ was calculated from a random 5% of the reflection data that was omitted from the subsequent refinement.

more widely. To help determine the most important features of Cmr7, we solved the structure of a second Cmr7 subunit, Sso1725. The sequence similarity between the two proteins is low (19% identity), and the crystal structure of Sso1725 (solved to 2.08 Å resolution; Table 1) shows limited structural similarity (Cα rmsd of 2.39 Å over 121 atoms for each monomer). The majority of the residues conserved in the six known orthologs of Cmr7 are located in the dimer interface and concave face (Figures 3C and S3). The electrostatic surfaces of the two proteins show broad similarities at the dimer interface, with symmetrical patches of negative charge at the poles and positive charge at the edges (Figure 3D). Given this level of conservation on only one face of the protein, we hypothesize that this region is important for function, possibly as a binding site for other CMR subunits. The stoichiometry of three dimers of Cmr7 per CMR complex is most easily accommodated if they form a trimeric (pseudo-hexamer) structure in the context of the complex.

## Electron Microscopy Reveals the Architecture of the CMR Complex

To elucidate the three-dimensional architecture of the CMR complex, we performed electron microscopy coupled to single-particle analysis experiments for the full complex and additionally for Cmr2/Cmr3/Cmr7 subcomplex, devoid of RNA (Figures 4 and S4). Projections of the maps matched with 2D averages assigned the same Euler angles (Figure S4), highlighting the validity of the maps. The resolution for both maps was ~25 Å, calculated at 0.5 Fourier Shell Correlation (3 sigma). The CMR complex exhibited cavities, compatible with an RNA threading machine. There was no obvious similarity to the "seahorse" structure of *E. coli* CASCADE (Jore et al., 2011). The Cmr2/Cmr3/Cmr7 subcomplex, which lacked bound crRNA, had overall dimensions of 90 × 120 × 110 Å, organized in a clamp or "crab claw" structure (Figure 4A). The intact CMR complex with loaded crRNA had overall dimensions of 160 × 120 × 110 Å with an upper "crab claw" connected to a protruding region (Figure 4B). These dimensions were compatible with the expected molecular masses of ~290 and ~430 kDa, respectively. The Cmr2/Cmr3/Cmr7 subcomplex fitted well to the upper region of intact CMR (Figure 4C), consistent with a role as

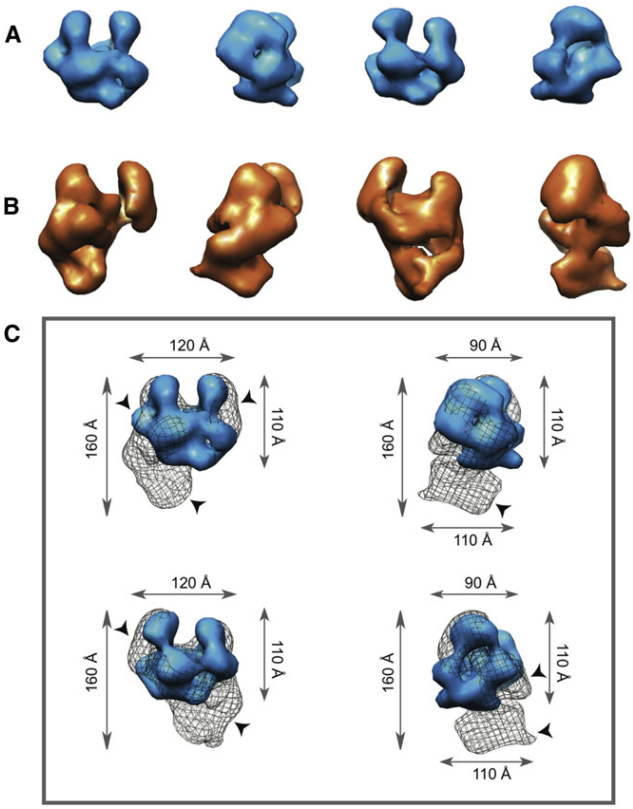



**Figure 4. 3D EM Visualization of CMR Complex**

(A) Surface representation of the Cmr2/Cmr3/Cmr7 subcomplex devoid of crRNA.

(B) Surface representation of the full CMR complex with bound crRNA.

(C) Superposition of Cmr2/Cmr3/Cmr7 (blue surface) on CMR/RNA (black mesh). Black arrowheads point to regions of additional density on the full CMR complex with bound crRNA compared to Cmr2/Cmr3/Cmr7. Gray arrows indicate dimensions in angstroms.

a scaffold for the assembly of the other Cmr subunits around the periphery (black arrows in Figure 4C).

## Ribonuclease Activity of SsoCMR

We tested the ability of SsoCMR to recognize and cleave RNA targets corresponding to spacers A1 and D63 in vitro. Both RNA targets were cleaved efficiently when a cognate crRNA (guide RNA) with an 8 nt 5′ tag was present (Figure 5). Manganese ions were essential for this activity, and magnesium could not substitute. ATP was not essential, but clearly stimulated the cleavage reaction (Figure S5). No crRNA-directed cleavage of DNA targets was observed (data not shown). To rule out the possibility of activity from a contaminating ribonuclease, the CMR complex was immunodepleted using antibodies raised against the Cmr7 subunit. Immunodepletion abolished the nuclease activity, suggesting strongly that the activity was associated with the CMR complex (Figure S5A). As a further control, native untagged SsoCMR purified from *S. solfataricus* cell extract by immunoprecipitation using the anti-Cmr7 antibody had the same activity as the column-purified, tagged protein complex (Figure S5B).

## Features of Guide and Target RNAs Important for Cleavage by CMR

The sequence and structural requirements of RNA cleavage by CMR were investigated by constructing a range of target and guide RNA molecules based on the spacer A1 and D63 sequences. To test for a molecular ruler mechanism as observed for PfuCMR, we reduced the length of the guide RNA (Figure 5A, panel ii). The cleavage sites did not move in register with the 3′ end of the guide RNA, suggesting that *Sulfolobus* and *Pyrococcus* CMR differ fundamentally in this respect. Deletion of the repeat-derived 5′ tag from the guide RNA abolished cleavage activity (Figure 5A, panel iii) and could not be rescued by substitution with a 5′ 8A sequence (panel iv), showing that this 5′ tag sequence is essential for cleavage, ruling out the possibility of a contaminating nuclease. The presence of an unpaired flap at the 3′ end of the target RNA was also required for activity (panel vi). This corresponds to the position of the PAM that is essential for cleavage of viral DNA targets by CASCADE (Gudbergsdottir et al., 2011; Lintner et al., 2011). However, for target RNA cleavage by CMR, the PAM sequence at this position was not essential, as a 6A sequence could substitute (Figure 5B, panel vii).

In addition to the target RNA, the guide RNA strand could also be cleaved by SsoCMR (Figure 5C, panel viii). Cleavage of the guide RNA was dependent on the presence of a 3′ overhang on the target RNA (Figure 5C, panel ix, xi). Guide and target were cleaved at approximately equal rates when present at an equimolar ratio, but at ratios of 20:1 or 5:1 excess of target over guide, the guide RNA was cleaved significantly more slowly. Under these conditions, multiple turnover cleavage of the target RNA was observed, suggesting that cleavage of the guide RNA was not essential for catalysis (Figure S5D).

## Sequence-Specific Cleavage by SsoCMR

The cleavage patterns observed for SsoCMR suggested a sequence- or structure-specific component to the activity. Sequence mapping suggested that strong cleavage always occurred at a UA dinucleotide in both the A1 and D63 target RNAs and the A1 guide RNA (Figure 5). Weaker cleavage was observed at UU dinucleotides. RNA cleavage by SsoCMR resulted in products with 3′-hydroxyl termini that could be extended by PolyA polymerase (Figure S6A). This is similar to the metal-dependent RNaseH-type activity observed for Piwi and Argonaute (Hutvagner and Simard, 2008). In contrast, the (metal-independent) Cas6 endonuclease yields 3′-cyclic phosphate products that are not extended by PolyA polymerase (Figure S6B). PfuCMR is also reported to generate 3′-cyclic phosphate products (Hale et al., 2009), another distinction between the two enzymes. To map the cleavage site of SsoCMR precisely, the cleavage at site 2 in the D63 target RNA was compared to a synthetic oligonucleotide terminating after the relevant UA dinucleotide (Figure S6C). The cleavage product generated by SsoCMR was 1 nt shorter than the oligonucleotide, consistent with cleavage at this position (and by extension at the other sites) as occurring at the center of the UA sequence.

To examine the importance of sequence for CMR-mediated cleavage of RNA, a D63-derived target RNA with only one UA site, corresponding to position 2, was synthesized. In the

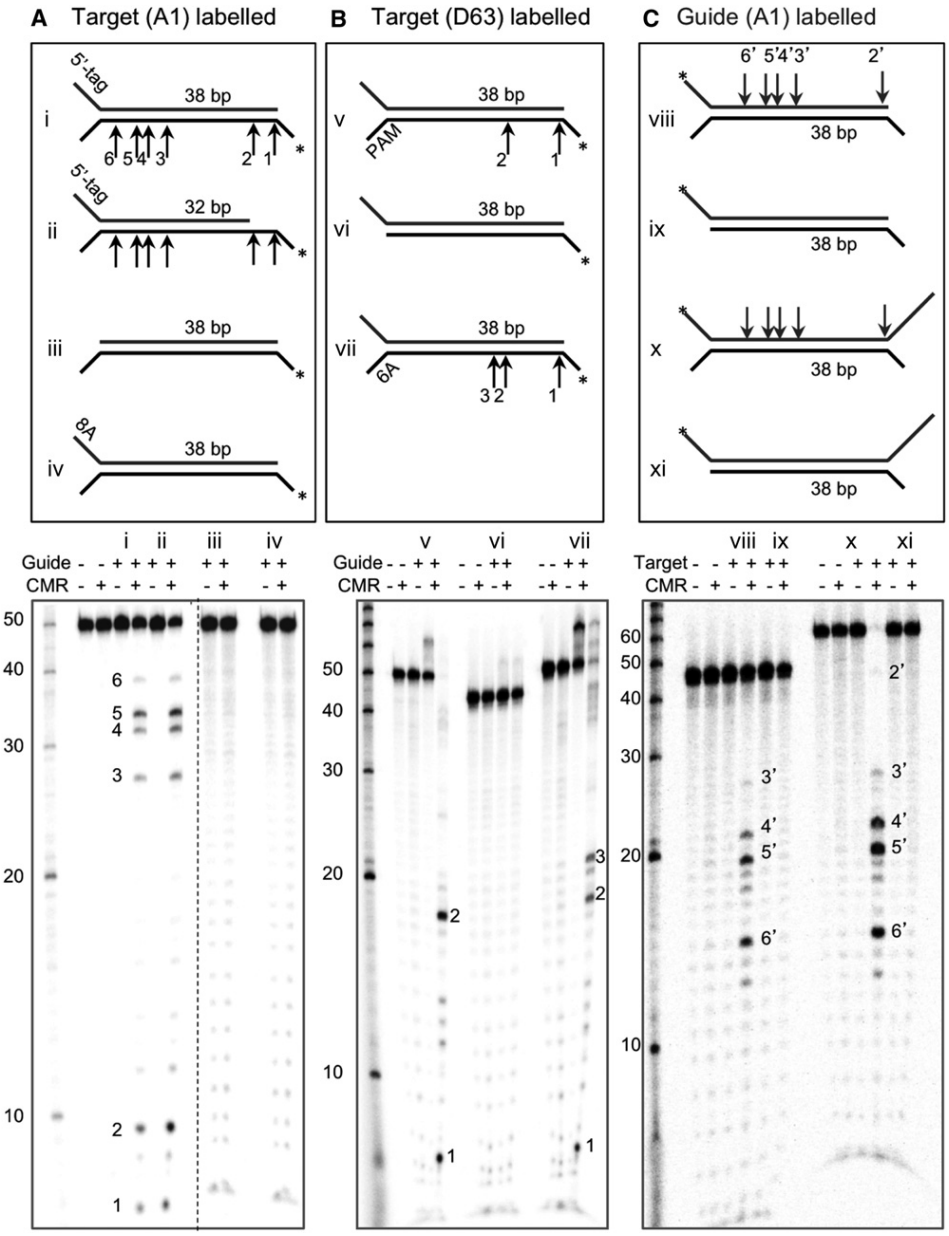



**Figure 5. Characterization of the *S. solfataricus* CMR Complex Activity In Vitro**

(A) Radiolabeled target RNA (5 μM) corresponding to spacer A1 was incubated with the Cmr7-tagged SsoCMR complex (0.5 μM) and guide RNA (1 μM). Cleavage of the target RNA was observed at six sites (labeled 1–6) (i). 3′ truncation of the guide RNA did not change the cleavage pattern of the target RNA, ruling out a molecular ruler mechanism (ii). The 8 nt 5′ tag was essential for cleavage activity, as either deletion (iii) or replacement with an 8A sequence (iv) abolished nuclease activity.

(B) The D63 target (0.5 μM) was cleaved in the presence of Cmr7-tagged SsoCMR complex (0.5 μM), cognate guide RNA (0.1 μM) (v). Deletion of the unpaired 3′ end of the target RNA, which corresponds to the position of the PAM sequence, abolished activity (vi). The presence of an unpaired 6A sequence at this position restored activity (vii).

(C) Radiolabeled guide RNA corresponding to spacer A1 (3 μM) was incubated with the Cmr7-tagged SsoCMR complex (0.5 μM) and cognate target RNA (1 μM). Cleavage was observed at up to five positions, labeled 2′ to 6′ (viii). This activity was dependent on the presence of the 3′ unpaired spacer sequence (ix) and was not influenced by the presence of an additional unpaired extension at the 3′ end of the guide RNA (x, xi). For each figure part, the 5′ end-labeled RNA strand is indicated with an asterisk. Labeled decade RNA markers (Ambion) are shown.

presence of the CMR complex and a cognate guide RNA sequence, this target was cleaved strongly at position 2, with only weak background cleavage elsewhere (Figure 6B). The same reaction product was observed when pairing this target with the wild-type guide, despite the presence of mismatches at the mutated UA sites. A modified version of the D63 guide

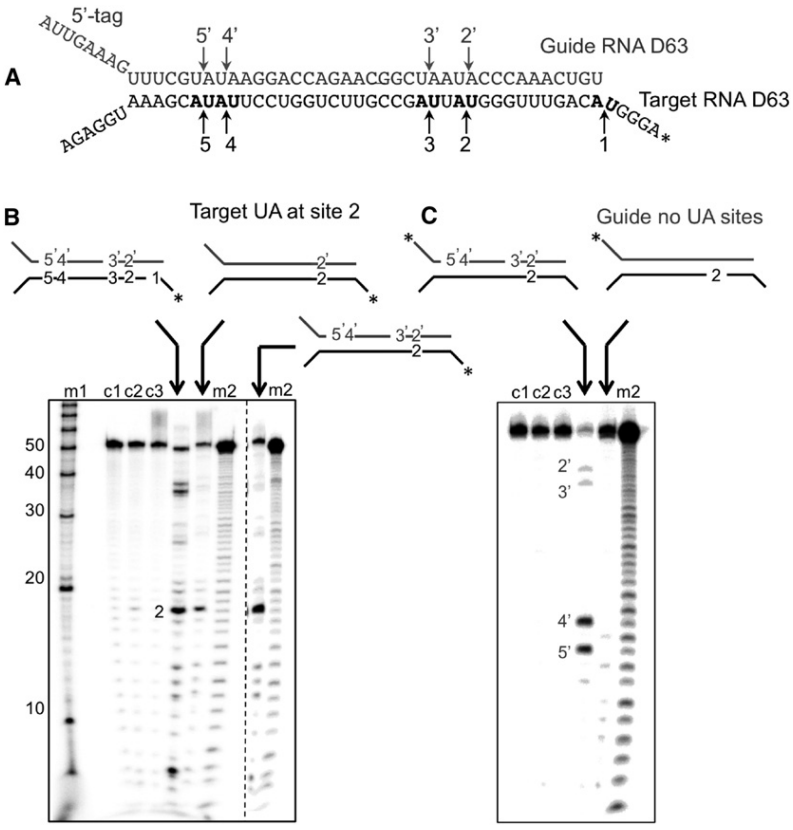



**Figure 6. The *S. solfataricus* CMR Complex Cleaves RNA Selectively at UA Sites**

(A) Sequence map of the D63 target and guide RNA.

(B) A D63 target oligonucleotide ($D63_{1UA}$, 0.5 μM) with UA cleavage sites 1, 3, 4, and 5 mutated to UG, was labeled and incubated with the Cmr3-tagged CMR complex (0.5 μM) and a cognate guide RNA ($crD63_{1UA}$, 0.05 μM). Cleavage of the target RNA was only observed at the single remaining UA site, site 2. The same target paired with the wild-type guide RNA gave identical cleavage products despite the presence of three mismatches.

(C) A D63 guide RNA sequence with all four UA sites mutated to CA ($crD63_{0UA}$, 0.5 μM) was labeled and incubated with the Cmr3-tagged CMR complex (0.5 μM) and a target RNA containing a single UA site at position 2 ($D63_{1UA}$, 0.05 μM). No significant cleavage of this target sequence was observed. The standard guide RNA (crD63) was cleaved at all four UA sites under the same conditions, despite the presence of mismatches at three positions in the RNA duplex. Control lanes for all gels are: m1, Ambion Decade markers; c1, labeled RNA alone; c2, labeled RNA and CMR; c3, both RNA strands without CMR. Asterisks indicate the 5′ RNA end labeled by $^{32}P$.

RNA lacking any UA sites was not cleaved by CMR in the presence of a target RNA with 1 UA site (Figure 6C). The presence of mismatches in the guide RNA opposite the target cleavage site did not abolish cleavage (Figure S6D), consistent with the mismatch tolerance observed previously for CASCADE (Gudbergsdottir et al., 2011; Semenova et al., 2011). Overall, these data confirmed that sequence-dependent cleavage of both the guide and target RNAs is a defining feature of SsoCMR and demonstrated that cleavage of the guide is not necessary for cleavage of the target, consistent with the observation of multiple turnover by the complex.

## DISCUSSION

### crRNA Content of SsoCMR

There are six CRISPR loci in *S. solfataricus* P2, named A–F (Lillestøl et al., 2006). Deep sequencing of 1.88 million crRNAs isolated from the CMR complex revealed a highly biased distribution, with 98% of the total crRNAs derived from locus D and C, an overrepresentation of 3.8- and 1.3-fold, respectively. The underrepresentation of crRNAs from highly transcribed loci A and B may be explained by differences in the processing of repeats by Cas6. It is notable that the CRISPR repeat sequence associated with the A and B loci is longer than those of C–F, which could provide a plausible basis for differential processing by Cas6. As *S. solfataricus* encodes five different Cas6 proteins, these may be specialized for the cleavage of particular classes of CRISPR repeat and/or may interact differently with the multiple CASCADE and CMR complexes in this organism. Another possibility is that differences in the removal of the repeat-derived sequence 3′ of the spacer by an unknown nuclease after Cas6 processing (Hale et al., 2009) influence incorporation efficiency.

The extreme variation in crRNA sequence numbers obtained from adjacent spacers in all the CRISPR arrays was also unexpected. In some cases, this is probably explained by the presence of internal promoter sequences encoded by adjacent spacers, driving higher transcript levels for particular regions of crRNA. Another variable that could influence crRNA processing is the potential for the formation of stable folded RNA structures following transcription that could influence crRNA processing by Cas6 and loading into the CMR complex. Each spacer has a unique sequence that confers a particular capacity to fold on the local crRNA structure. To address this further, we analyzed the thermodynamic stability of folded RNA structures for each crRNA derived from CRISPR locus C and plotted these against the frequency of occurrence of each crRNA in the CMR sequencing sample (Figure S2). The crRNAs with the potential to fold into the most stable structures were clearly not highly represented in the data set, while the most highly represented sequences had similar, modest folding propensity. This is suggestive of an influence of the secondary structure of the crRNA locally on the efficiency of cleavage by Cas6 and possibly on the loading of crRNA into the CMR complex. Such effects may be ameliorated in the thermophiles by the high growth temperature, but could constitute a significant problem in temperature mesophiles. In organisms such as *E. coli*, the repeat sequence is palindromic, folding into a stable hairpin structure that is recognized by the CasE nuclease (Brouns et al., 2008). This may be an evolutionary mechanism that helps impose an

ordered secondary structure on the crRNA to avoid problems due to folding of spacer sequences. In contrast, Cas6, which is present in many thermophiles, binds unstructured crRNA repeats (Wang et al., 2011).

## Structure of the CMR Complex

The EM envelope of the Cmr2/Cmr3/Cmr7 subcomplex bears some similarity to the clamp region of RNA polymerase. In particular, the distance between the two sides of the "claw" is ~30 Å wide, and their length is ~40 Å. It is tempting to view this feature as a dsRNA binding cleft, particularly as Cmr2 is assumed to harbor the active site of the CMR complex. In support of this, the cleft is somewhat deeper in the Cmr2/Cmr3/Cmr7 subcomplex, which contains no bound crRNA, in comparison to the full complex. The lack of crRNA in the Cmr2/Cmr3/Cmr7 subcomplex fits with a presumed role for the RAMP-containing Cmr subunits (Cmr1, Cmr4, and Cmr6) in RNA binding (Makarova et al., 2011a). This is consistent with the recent prediction that Cmr3 is most closely related to the Cas5 subunit of CASCADE, while Cmr1, Cmr4, and Cmr6 are closer matches to the Cas7 subunit (CasC in *E. coli*), which is known to bind crRNA (Makarova et al., 2011a). Compared to Cmr2/Cmr3/Cmr7, the additional Cmr subunits are distributed mainly at the front and at the bottom of the complete CMR complex and may form a structure related to the crRNA binding CasC backbone of CASCADE. Visualization of the path of RNA in the CMR structure remains a key aim to help elucidate the mechanism and molecular organization.

## The Mechanism of SsoCMR in Viral RNA Cleavage

We have demonstrated that the SsoCMR complex cleaves target RNA in a sequence-specific manner that is dependent on the presence of a guide crRNA with a cognate sequence and an 8 nt repeat-derived tag at the 5′ end. Cleavage occurs at UA dinucleotides, generating products with 3′-OH and 5′-phosphate ends, like those produced by Argonaute and Piwi (Hutvagner and Simard, 2008). The cleavage sequence is extremely common in the *Sulfolobales* and their viruses. Of the >400 spacers in the CRISPR loci of the *S. solfataricus* P2 genome, only 11 lack a UA dinucleotide. The palindromic nature of this dinucleotide ensures that an equivalent UA sequence is always present in the guide RNA. Cleavage of both the target and guide RNA at UA sites was observed, with the guide cleaved significantly more slowly at high ratios of target:substrate. Under these conditions, one guide RNA molecule could support the cleavage of several target RNAs, demonstrating that cleavage of the guide is not essential for target RNA destruction.

Although not yet shown directly, the N-terminal permuted HD nuclease domain present in Cmr2 is generally assumed to be the nuclease site of the CMR complex. The distantly related HD domain present in Cas3 has been shown capable of cleaving both RNA and DNA, although they are specific for single-stranded nucleic acids and generate products with 3′-cyclic phosphates (Mulepati and Bailey, 2011; Beloglazova et al., 2011). It is also possible that the RNA cleavage activity resides elsewhere in the Cmr2 subunit or in one or more of the RAMP-containing subunits (Cmr1, Cmr2, Cmr4, and Cmr6), which are distantly related to Cas5, Cas6, and Cas7 (Makarova et al., 2011a).

We observed no requirement for a PAM sequence adjacent to the protospacer in RNA targets. This is a marked difference from the DNA targeting CASCADE in archaea, which only cleaves substrates containing a PAM (Gudbergsdottir et al., 2011; Manica et al., 2011). The primary role of the PAM may be to maintain a mismatched region between the 5′ tag of the crRNA and the sequence immediately adjacent to the spacer of the target. For DNA targeting systems, this ensures that the chromosomal CRISPR locus is not targeted for cleavage (Marraffini and Sontheimer, 2010b; Mojica et al., 2009). Given that the CMR complex only targets viral RNA, there is no requirement for PAM detection to operate in this case, as the host genome will not be a target. Although no PAM is required, CMR-mediated cleavage requires an unpaired RNA region at the 3′ end of the target RNA, downstream of the protospacer. This discrimination may be at the structural rather than sequence level and is consistent with a role in vivo in targeting viral mRNA sequences, which will typically be considerably longer than the guide RNA species.

The observed activity of SsoCMR differs markedly from that reported previously for the enzyme from *P. furiosus* (Hale et al., 2009). PfuCMR operates by a molecular ruler mechanism without sequence-dependent cleavage, generates 3′-cyclic phosphate products, and does not require an extension at the 3′ end of the target RNA, at least in vitro. The long evolutionary distance between the two species may explain these differences. CMR complexes have been classified into five families (A–E) on the basis of the sequence of the large subunit, Cmr2 (Garrett et al., 2011). On this basis, the SsoCMR complex belongs to family B while PfuCMR is one of the very few representatives of family C. To put these differences in context, the type IIIA, CMR-like complex from *Staphylococcus epidermidis* targets DNA rather than RNA (Marraffini and Sontheimer, 2008). This may therefore reflect the plasticity inherent in the CRISPR system.

The RNA targeting functionality of the CMR complex in prokaryotes has parallels with the eukaryal piRNA pathway that uses guide RNA to recognize and cleave the mRNA of mobile genetic elements (Aravin et al., 2007). As in the CRISPR system, in the piRNA pathway, the small guide RNAs are generated by cleavage of a long mRNA transcript, loaded into an endoribonuclease (Piwi), and used to target and degrade the mobile mRNA by means of dsRNA cleavage, yielding products with 3′-OH termini (reviewed in Nowotny and Yang, 2009). There are, though, important differences. Piwi recognizes the 5′ phosphate of the guide RNA specifically (Ma et al., 2005; Parker et al., 2005), while the guide RNA generated by Cas6 and utilized by CMR lacks a 5′ phosphate and has an essential 8 nt 5′ tag. More fundamentally, there is no obvious homology between the Piwi and CMR proteins. Piwi uses an RNaseH domain for dsRNA cleavage, which is absent from the CMR complex. The stimulation of CMR cleavage activity by ATP suggests that an ATP-driven conformational change may be utilized to reposition the dsRNA with respect to the active site. In the absence of any detectable Walker A or B motifs, the likely site for ATP binding is the C-terminal polymerase/cyclase domain of Cmr2. This domain has no known function but is expected to bind nucleotide triphosphates. If such an ATP-dependent RNA

repositioning mechanism were in operation, it would constitute another aspect of the CMR complex.

## Concluding Remarks

In summary, we provide a low-resolution structure of the CMR complex for prokaryotic viral RNA degradation. Deep sequencing suggests that posttranscriptional processing may exert considerable influence on the loading of its crRNA component. The reaction mechanism involves manganese-dependent and ATP-stimulated ribonuclease activity that degrades both target and guide RNA in a sequence-dependent manner. Future studies will aim in particular to map the individual subunits within the EM envelope and the course of the bound crRNA in the complex and to define the function of the polymerase/cyclase domain and the role of ATP in the reaction.

## EXPERIMENTAL PROCEDURES

### Cloning, Expression, and Purification of Cmr7 Paralogs

Details of cloning, purification, and crystal structure solution for Sso1986 were reported previously (Oke et al., 2010). Sso1725 was expressed and purified according to published protocols (Oke et al., 2010). Briefly, full-length *sso*1725 was cloned into the pDEST14 vector with an N-terminal 6xHis tag and overexpressed in C43 (DE3) *E. coli* at 37°C in LB medium. Expression was induced using 0.4 mM IPTG, and the cultures were harvested after overnight incubation at 25°C. The cell pellets were resuspended and lysed (Soniprep 150, MSE). The lysate was clarified by centrifugation, and then protein was purified by immobilized nickel affinity chromatography and size-exclusion chromatography (Superdex 75 column, GE Healthcare). Sso1725 was concentrated to 10 $mg.ml^{-1}$ for crystallization.

### Antibody Generation

Sheep polyclonal antibodies were raised against the recombinant Cmr7 (Sso1986) protein and supplied by the Scottish National Blood Transfusion Service, Pentlands Science Park, Midlothian.

### RNA Oligonucleotides Used for CMR Activity Assays

RNA oligonucleotides were chemically synthesized (Integrated DNA Technologies), end labeled with $^{32}$P-ATP, and purified by denaturing gel electrophoresis. The sequences used are listed in the Supplemental Experimental Procedures.

### Purification of the Native CMR Complex from *S. solfataricus*

*S. solfataricus* strain P2 biomass was grown as described previously (Götz et al., 2007). The CMR complex was purified over four column chromatography steps, and purification was followed using an antibody raised against subunit Cmr7, as described in the Supplemental Experimental Procedures. This yielded the homogeneous complex shown in Figure 1.

### Expression and Purification of Tagged CMR Complex in *S. solfataricus*

This was carried out as described previously by cloning the relevant gene into entry vector pMZ1 (Zolghadr et al., 2007), followed by subcloning into expression vector pSVA9, expressing the relevant subunit with a C-terminal strep-His tag (Albers et al., 2006), described in detail in the Supplemental Experimental Procedures. The Cmr2/Cmr3/Cmr7 subcomplex analyzed by electron microscopy was obtained during purification of the CMR complex with a tagged Cmr3 subunit. The subcomplex eluted separately from the full complex on gel filtration and contained no bound crRNA (Figure S5).

### RNA Isolation and Sequencing

RNA was extracted from the purified native CMR complex by the classical phenol/chloroform method followed by ethanol precipitation and vacuum desiccation. Dried RNA was resuspended in 5 μl water and directly labeled in a 10 μl reaction containing polynucleotide kinase and 2 μCi $\gamma^{32}$P-ATP. Labeled RNAs were analyzed by electrophoresis on a 15% acrylamide, 7 M urea, TBE denaturing gel and visualized by phosphorimaging. For crRNA deep sequencing, small RNA sequences were generated by the GenePool at the University of Edinburgh using the Illumina small RNA prep kit v1 and subjected to high-throughput sequencing using a Genome Analyzer IIx. This resulted in the addition of the adaptor sequence TCGTATGCCGTCTTCTGCTTG at the 3′ end of each sequence. The adaptor sequence was trimmed away from the reads with a bespoke Perl script. Reads were mapped against the *S. solfataricus* P2 genome with BWA (Li and Durbin, 2009) using default parameters and converted into BAM using SAMtools (Li et al., 2009). Of the 2,527,217 reads, 1,997,151 were mapped (79%). The number and strand orientation of the reads mapping to each spacer were quantified. The raw data from the sequencing run is available from the corresponding author on request, and the sequences mapping onto each spacer are listed in Table S1. The raw sequence data have been uploaded to the Sequence Read Archive with accession number ERP001053 (http://www.ebi.ac.uk/ena/data/view/ERP001053).

### RNA Cleavage Assays

Purified SsoCMR complex and unlabeled guide RNA were mixed in buffer (20 mM Mes·HCl [pH 6.0], 100 mM potassium glutamate, 10 mM DTT, 10 mM $MnCl_2$, and the RNase inhibitor SUPERase.In [Ambion]) and preincubated at room temperature for 10 min prior to the addition of 5′-$^{32}$P-end-labeled synthetic target RNA to the reaction mix. Target and guide RNA and CMR complex concentrations are indicated in the figure legends. The reaction was further incubated at 75°C for 10 min in standard assays or for the time indicated in the figure. Reactions were stopped by chilling on ice and addition of formamide loading buffer. Samples were separated on 20% polyacrylamide, 7 M urea, 1× TBE gels. Electrophoresis was completed at 90 W, 50°C for 90 min, and the gels were visualized by phosphorimaging. 5′ end-labeled RNA size standards (Decade Markers, Ambion) were used to determine the sizes of the observed products. Cas6 activity was assayed as described previously (Lintner et al., 2011).

### Crystallography of Sso1725

Sso1725 crystals were grown at 20°C using the sitting drop vapor diffusion method. A reservoir of 0.15 M sodium acetate (pH 5.6), 2.0 M ammonium sulfate was used, and protein was mixed with the precipitant at a ratio of 2:1. Crystals were cryoprotected by successively soaking in solutions of reservoir containing 8%, 16%, 20%, and 25% glycerol before freezing in liquid nitrogen. A native data set at 2.08 Å resolution was collected at Diamond Light Source (Beamline I03). An anomalous SAD data set was collected at the same beamline using a native crystal briefly soaked in reservoir solution containing 42 mM samarium chloride before cryoprotecting as described above. Data sets were processed and refined using the methods described in the Supplemental Experimental Procedures. The coordinates were deposited in the PDB under the accession code 2XVO.

### EM Studies

The intact CMR complex bound to crRNA and the Cmr2/Cmr3/Cmr7 subcomplex were studied by negative-staining electron microscopy and single-particle analysis. Data were collected on an FEI F20 FEG microscope, equipped with an 8 k × 8 k CCD camera. Images were collected under low-dose mode at a magnification of 50,000×, at a final sampling of 1.6 Å/pixel at the specimen level. Single-particle images were selected interactively using the Boxer program from the EMAN single-particle analysis package (Ludtke et al., 1999) and extracted into boxes. Image processing was performed using the IMAGIC-5 package (van Heel et al., 1996). The data set was resampled at 6.4 Å/pixel. 10,235 (CMR/RNA) and 5,612 (Cmr2/Cmr3/Cmr7) images were band-pass filtered with a high pass cutoff of 110 Å and a low pass cutoff of 18 Å. The single-particle images were analyzed by Multivariate Statistical Analysis with IMAGIC-5. The data set was subjected to successive rounds of alignment and classification in order to improve the resulting image class averages. Selected CMR/RNA class averages were used to calculate a starting 3D volume by common lines using the Euler program in the IMAGIC-5 package. The CMR/RNA structure was refined until the map converged.

We used the CMR/RNA map to align Cmr2/Cmr3/Cmr7 images and to assign Euler angles by projection matching. Subsequent refinement was carried out until the Cmr2/Cmr3/Cmr7 map converged. Figures were prepared with UCSF Chimera (Goddard et al., 2007).

## ACCESSION NUMBERS

The crystal structure of the Cmr7 subunit of CMR (Sso1725) has been submitted to the Protein Data Bank with accession number 2X5Q. The deep sequencing data for the RNA bound to the CMR complex has been uploaded to the Sequence Read Archive with accession number ERP001053.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures, two tables, and Supplemental Experimental Procedures and can be found with this article online at doi:10.1016/j.molcel.2011.12.013.

## ACKNOWLEDGMENTS

Thanks to Paul Talbot and the University of St. Andrews Mass Spectrometry service for expert technical assistance. This project was funded by the Biotechnology and Biological Sciences Research Council grant number BB/G011400/1. The Electron Microscopy Facility at Edinburgh is supported by the Scottish Alliance for Life Sciences and the Wellcome Trust (WT087658MA). J. Reimann and S.-V.A. were supported by intramural funds of the Max Planck Society.

Received: August 10, 2011
Revised: November 15, 2011
Accepted: December 5, 2011
Published online: January 5, 2012

## REFERENCES

Albers, S.V., Jonuscheit, M., Dinkelaker, S., Urich, T., Kletzin, A., Tampé, R., Driessen, A.J., and Schleper, C. (2006). Production of recombinant and tagged proteins in the hyperthermophilic archaeon *Sulfolobus solfataricus*. Appl. Environ. Microbiol. *72*, 102–111.

Aravin, A.A., Hannon, G.J., and Brennecke, J. (2007). The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science *318*, 761–764.

Beloglazova, N., Petit, P., Flick, R., Brown, G., Savchenko, A., and Yakunin, A.F. (2011). Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference. EMBO J. *30*, 4616–4627.

Bolotin, A., Quinquis, B., Sorokin, A., and Ehrlich, S.D. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology *151*, 2551–2561.

Brouns, S.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J., Snijders, A.P., Dickman, M.J., Makarova, K.S., Koonin, E.V., and van der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. Science *321*, 960–964.

Carte, J., Wang, R., Li, H., Terns, R.M., and Terns, M.P. (2008). Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. Genes Dev. *22*, 3489–3496.

Deltcheva, E., Chylinski, K., Sharma, C.M., Gonzales, K., Chao, Y., Pirzada, Z.A., Eckert, M.R., Vogel, J., and 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, J.E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P., and Moineau, S. (2008). Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. J. Bacteriol. *190*, 1390–1400.

Garrett, R.A., Shah, S.A., Vestergaard, G., Deng, L., Gudbergsdottir, S., Kenchappa, C.S., Erdmann, S., and She, Q. (2011). CRISPR-based immune systems of the Sulfolobales: complexity and diversity. Biochem. Soc. Trans. *39*, 51–57.

Goddard, T.D., Huang, C.C., and Ferrin, T.E. (2007). Visualizing density maps with UCSF Chimera. J. Struct. Biol. *157*, 281–287.

Götz, D., Paytubi, S., Munro, S., Lundgren, M., Bernander, R., and White, M.F. (2007). Responses of hyperthermophilic crenarchaea to UV irradiation. Genome Biol. *8*, R220.

Gudbergsdottir, S., Deng, L., Chen, Z., Jensen, J.V., Jensen, L.R., She, Q., and Garrett, R.A. (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.

Haft, D.H., Selengut, J., Mongodin, E.F., and Nelson, K.E. (2005). A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput. Biol. *1*, e60.

Hale, C.R., Zhao, P., Olson, S., Duff, M.O., Graveley, B.R., Wells, L., Terns, R.M., and Terns, M.P. (2009). RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell *139*, 945–956.

Horvath, P., and Barrangou, R. (2010). CRISPR/Cas, the immune system of bacteria and archaea. Science *327*, 167–170.

Horvath, P., Romero, D.A., Coûté-Monvoisin, A.C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C., and Barrangou, R. (2008). Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. J. Bacteriol. *190*, 1401–1412.

Hutvagner, G., and Simard, M.J. (2008). Argonaute proteins: key players in RNA silencing. Nat. Rev. Mol. Cell Biol. *9*, 22–32.

Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, U., Wurm, R., Wagner, R., et al. (2011). Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat. Struct. Mol. Biol. *18*, 529–536.

Karginov, F.V., and Hannon, G.J. (2010). The CRISPR system: small RNA-guided defense in bacteria and archaea. Mol. Cell *37*, 7–19.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lillestøl, R.K., Redder, P., Garrett, R.A., and Brügger, K. (2006). A putative viral defence mechanism in archaeal cells. Archaea *2*, 59–72.

Lillestøl, R.K., Shah, S.A., Brügger, K., Redder, P., Phan, H., Christiansen, J., and Garrett, R.A. (2009). CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties. Mol. Microbiol. *72*, 259–272.

Lintner, N.G., Kerou, M., Brumfield, S.K., Graham, S., Liu, H., Naismith, J.H., Sdano, M., Peng, N., She, Q., Copié, V., et al. (2011). Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). J. Biol. Chem. *286*, 21643–21656.

Ludtke, S.J., Baldwin, P.R., and Chiu, W. (1999). EMAN: semiautomated software for high-resolution single-particle reconstructions. J. Struct. Biol. *128*, 82–97.

Ma, J.B., Yuan, Y.R., Meister, G., Pei, Y., Tuschl, T., and Patel, D.J. (2005). Structural basis for 5′-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. Nature *434*, 666–670.

Makarova, K.S., Aravind, L., Wolf, Y.I., and Koonin, E.V. (2011a). Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. Biol. Direct *6*, 38.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J., Wolf, Y.I., Yakunin, A.F., et al. (2011b). Evolution and classification of the CRISPR-Cas systems. Nat. Rev. Microbiol. *9*, 467–477.

Manica, A., Zebec, Z., Teichmann, D., and Schleper, C. (2011). In vivo activity of CRISPR-mediated virus defence in a hyperthermophilic archaeon. Mol. Microbiol. *80*, 481–491.

Marraffini, L.A., and Sontheimer, E.J. (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science *322*, 1843–1845.

Marraffini, L.A., and Sontheimer, E.J. (2010a). CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat. Rev. Genet. *11*, 181–190.

Marraffini, L.A., and Sontheimer, E.J. (2010b). Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature *463*, 568–571.

Mojica, F.J., Díez-Villaseñor, C., García-Martínez, J., and Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology *155*, 733–740.

Mulepati, S., and Bailey, S. (2011). Structural and biochemical analysis of nuclease domain of clustered regularly interspaced short palindromic repeat (CRISPR)-associated protein 3 (Cas3). J. Biol. Chem. *286*, 31896–31903.

Nowotny, M., and Yang, W. (2009). Structural and functional modules in RNA interference. Curr. Opin. Struct. Biol. *19*, 286–293.

Oke, M., Carter, L.G., Johnson, K.A., Liu, H., McMahon, S.A., Yan, X., Kerou, M., Weikart, N.D., Kadi, N., Sheikh, M.A., et al. (2010). The Scottish Structural Proteomics Facility: targets, methods and outputs. J. Struct. Funct. Genomics *11*, 167–180.

Parker, J.S., Roe, S.M., and Barford, D. (2005). Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex. Nature *434*, 663–666.

Semenova, E., Jore, M.M., Datsenko, K.A., Semenova, A., Westra, E.R., Wanner, B., van der Oost, J., Brouns, S.J., and Severinov, K. (2011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc. Natl. Acad. Sci. USA *108*, 10098–10103.

Shah, S.A., Hansen, N.R., and Garrett, R.A. (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.

van Heel, M., Harauz, G., Orlova, E.V., Schmidt, R., and Schatz, M. (1996). A new generation of the IMAGIC image processing system. J. Struct. Biol. *116*, 17–24.

Wang, R., Preamplume, G., Terns, M.P., Terns, R.M., and Li, H. (2011). Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage. Structure *19*, 257–264.

Wurtzel, O., Sapra, R., Chen, F., Zhu, Y., Simmons, B.A., and Sorek, R. (2010). A single-base resolution map of an archaeal transcriptome. Genome Res. *20*, 133–141.

Zolghadr, B., Weber, S., Szabó, Z., Driessen, A.J., and Albers, S.V. (2007). Identification of a system required for the functional surface localization of sugar binding proteins with class III signal peptides in *Sulfolobus solfataricus*. Mol. Microbiol. *64*, 795–806.

Molecular Cell
Article

# Structure of the CRISPR Interference Complex CSM Reveals Key Similarities with Cascade

Christophe Rouillon,[1,4] Min Zhou,[3,4] Jing Zhang,[1] Argyris Politis,[3] Victoria Beilsten-Edmands,[3] Giuseppe Cannone,[2] Shirley Graham,[1] Carol V. Robinson,[3,*] Laura Spagnolo,[2,*] and Malcolm F. White[1,*]
[1]Biomedical Sciences Research Complex, University of St Andrews, Fife KY16 9ST, UK
[2]Institute of Structural Molecular Biology and Centre for Science at Extreme Conditions, University of Edinburgh, Edinburgh EH9 3JR, UK
[3]Department of Chemistry, 12 Mansfield Road, University of Oxford, Oxford OX1 3TA, UK
[4]These authors contributed equally to this work
*Correspondence: carol.robinson@chem.ox.ac.uk (C.V.R.), laura.spagnolo@ed.ac.uk (L.S.), mfw2@st-andrews.ac.uk (M.F.W.)
http://dx.doi.org/10.1016/j.molcel.2013.08.020


## SUMMARY

**The Clustered Regularly Interspaced Palindromic Repeats (CRISPR) system is an adaptive immune system in prokaryotes. Interference complexes encoded by CRISPR-associated (*cas*) genes utilize small RNAs for homology-directed detection and subsequent degradation of invading genetic elements, and they have been classified into three main types (I–III). Type III complexes share the Cas10 subunit but are subclassifed as type IIIA (CSM) and type IIIB (CMR), depending on their specificity for DNA or RNA targets, respectively. The role of CSM in limiting the spread of conjugative plasmids in *Staphylococcus epidermidis* was first described in 2008. Here, we report a detailed investigation of the composition and structure of the CSM complex from the archaeon *Sulfolobus solfataricus*, using a combination of electron microscopy, mass spectrometry, and deep sequencing. This reveals a three-dimensional model for the CSM complex that includes a helical component strikingly reminiscent of the backbone structure of the type I (Cascade) family.**

## INTRODUCTION

The Clustered Regularly Interspaced Palindromic Repeats (CRISPR) system is a prokaryotic adaptive immune system that targets and degrades invading genetic elements. DNA fragments from mobile elements are captured and incorporated into the host genome at a CRISPR locus, flanked by direct repeat sequences, in a poorly understood process termed ''adaptation'' (van der Oost et al., 2009; Yosef et al., 2012). Transcription of the locus generates a long pre-CRISPR RNA (pre-crRNA) transcript that is processed into unit-length crRNAs by specific cleavage. Each crRNA is composed of a single ''spacer'' region homologous to a mobile genetic element, with a variable flanking region derived from the CRISPR sequence that flanks the spacer. crRNAs are loaded into a ribonucleoprotein complex and utilized for homology-dependent targeting and cleavage of cognate mobile elements in a process known as ''interference'' (Marraffini and Sontheimer, 2008). These complexes have been classified into three major types, I–III, characterized by the presence of a signature CRISPR-associated (Cas) protein: Cas3, Cas9, and Cas10 for types I, II, and III, respectively (Makarova et al., 2011b). In addition, types I and III share a variable number of Repeat Associated Mysterious Protein (RAMP) subunits. The RAMP domain is a derivative of the RNA Recognition Motif (RRM) fold and is often involved in RNA binding and/or cleavage (Makarova et al., 2011a).

The type IIIA complex, also known as the CSM complex, is found in a wide variety of bacteria and archaea. In *Staphylococcus epidermidis*, CSM is encoded in an operon that includes the *csm1-6* genes and has been shown to limit plasmid conjugation by targeting invading DNA for degradation (Marraffini and Sontheimer, 2008). CSM is associated with crRNA generated by cleavage of pre-crRNA by Cas6 and 3′-end processing by an unknown nuclease (Hatoum-Aslan et al., 2011). The CRISPR locus in the host genome is not cleaved by type IIIA systems, as there is a requirement for a mismatch region at the boundary of the repeat-spacer sequence: a condition that is met for foreign DNA targets but not for the genomic locus, where the crRNA matches perfectly to the genomic sequence (Marraffini and Sontheimer, 2010).

Although the type IIIA systems provided the first example of unequivocal DNA targeting by the CRISPR system, there has been little progress in the biochemical characterization of any CSM complex. Here, we report the purification and structural characterization of the CSM complex from the archaeon *Sulfolobus solfataricus*. Electron microscopy (EM) reveals an extended, intertwined helical conformation that suggests a backbone formed by RAMP subunits with striking similarities to that of the type IE Cascade complex (Wiedenheft et al., 2011). Mass spectrometry (MS) was used to define the subunit composition and subcomplex organization. Deep sequencing of the crRNA copurifying with the complex unveils a remarkable specificity for crRNA that suggests a very biased uptake mechanism, perhaps coupled to the Cas6 endonuclease.

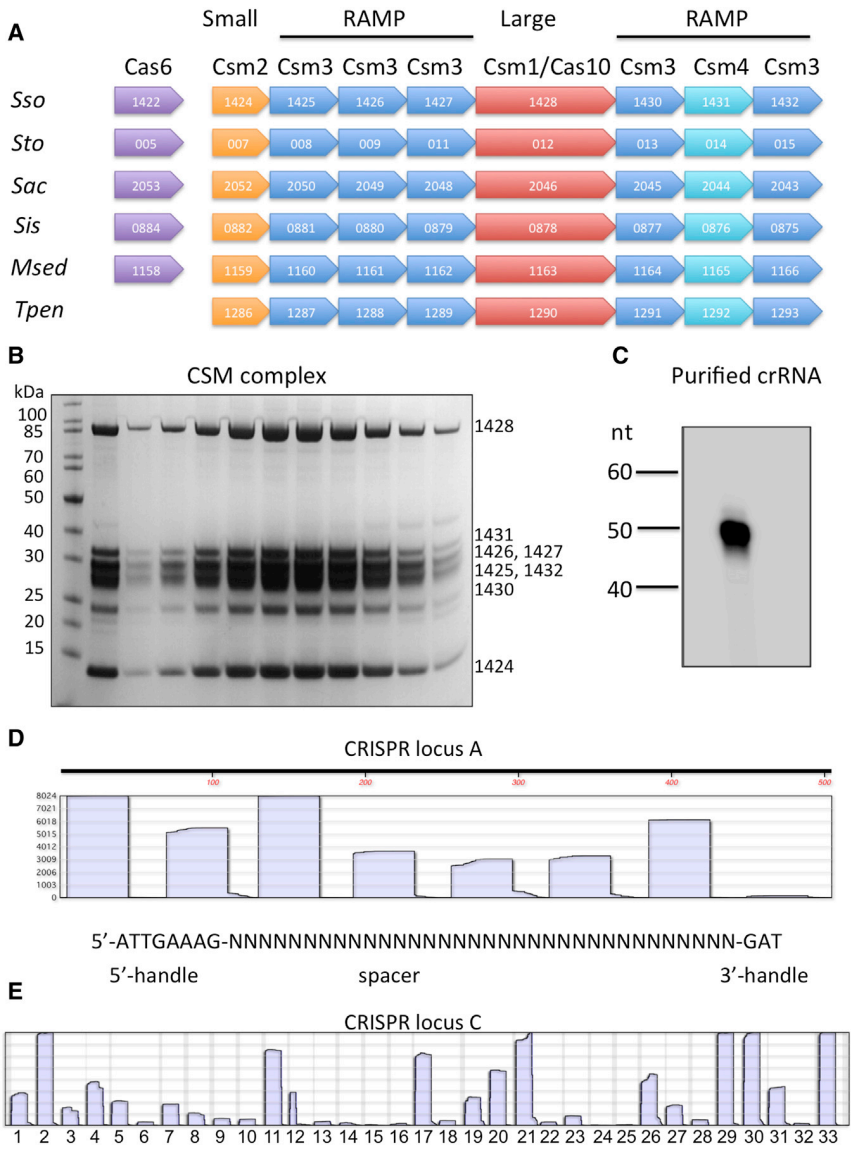



**Figure 1. Purification and RNA Content of the CSM Complex from *S. solfataricus***

(A) Gene organization of the *csm* locus in selected crenarchaeal species. The gene order is conserved and typically includes a gene encoding Cas6 for crRNA processing. Gene numbers are shown and are contiguous on the genome. Abbreviations are as follows: Sso, *S. solfataricus*; Sto, *S. tokodaii*; Sac, *S. acidocaldarius*; Sis, *S. islandicus* strain M.14.25; Msed, *Metallosphaera sedula*; Tpen, *Thermofilum pendens*. *T. pendens* Cas6 is present elsewhere on the genome.

(B) Fractions of the CSM complex eluting from the final gel-filtration column during purification. All eight subunits can be visualized and detected by MS.

(C) RNA purified from the purified CSM complex. A single discrete band around 50 nt was observed.

(D) A linear coverage map for a series of eight spacers from the *S. solfataricus* P1 A locus is shown. 5′ ends were defined by the 8 nt 5′-handle derived from cleavage of the repeat by Cas6.

(E) Linear coverage map for the entire CRISPR C locus from *S. solfataricus* P1, highlighting the variability in coverage.

## RESULTS

### Identification and Purification of a CSM Complex from *S. solfataricus*

The organization of *csm* loci in a selection of crenarchaea (Figure 1A) shows conservation of gene order across these species, with eight *csm* genes typically preceded by a *cas6* gene for crRNA processing. In addition to RAMP-domain subunits, partially conserved "large" and "small" subunits have been identified in most type I and III interference complexes (Makarova et al., 2011a). The gene encoding the large subunit Csm1 is annotated as a split gene encompassing *sso1428* and *sso1429* in the *S. solfataricus* genome sequence (She et al., 2001) but was found intact in the current study and named *sso1428*. The small subunit, Sso1424, probably corresponds to Csm2 in *S. epidermidis*, although sequence similarity is very limited. There are six *csm* genes encoding RAMP domain proteins (Figure 1A), compared to three in *S. epidermidis* (*csm3–csm5*) (Marraffini and Sontheimer, 2008). Previously, Sso1431 was predicted to be a member of the Csm4 family (Makarova et al., 2011a), and sequence comparison using HHPRED (Söding et al., 2005) reveals a clear structural match to the *P. furiosus* Cmr3 subunit of the type IIIB CMR complex for RNA targeting (Shao et al., 2013), in agreement with the prediction that both Csm4 and Cmr3 are members of the Cas5 subfamily of RAMP proteins (Makarova et al., 2011a). Of the other RAMP subunits, Sso1426 has the closest predicted structural match to the Cas7 family (Lintner et al., 2011), which is thought to make up the crRNA binding helical backbone of all the type I complexes (Makarova et al., 2011a; van Duijn et al., 2012). We therefore assign this subunit to the Csm3 family as suggested previously (Makarova et al., 2011a). Other RAMP subunits are considered as Csm3 paralogs in Figure 1.

We purified the CSM complex from *S. solfataricus* by employing an approach previously used to isolate the type IIIB (CMR) complex (Zhang et al., 2012). This involved expression of one tagged subunit of the complex, in this case either the *sso1428* or *sso1431* gene encoding Csm1 or Csm4, respectively, from a viral expression vector, followed by a combination of affinity, ion exchange, and gel filtration chromatography. The complex was purified as a homogeneous population eluting as a single peak in the final chromatography step, as confirmed by SDS-PAGE analysis (Figure 1B). The presence of all eight subunits

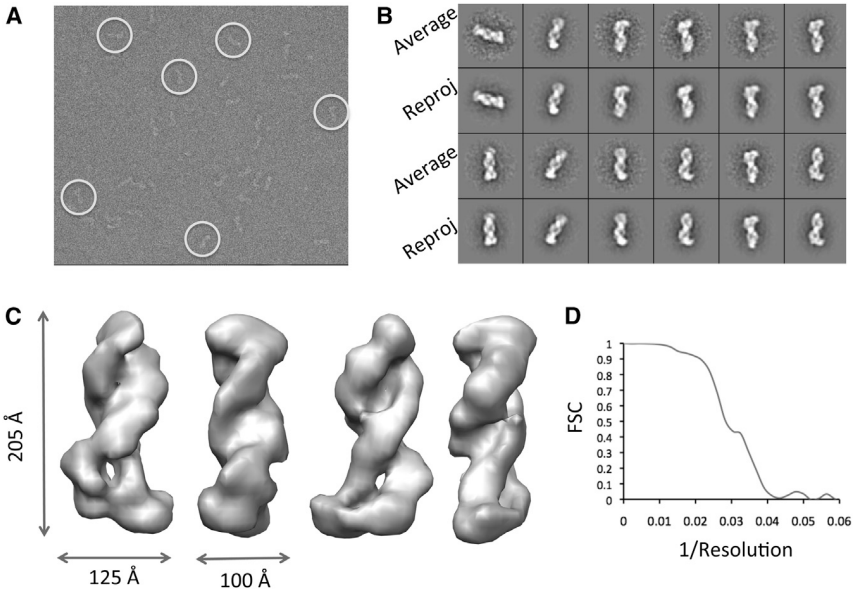



**Figure 2. 3D-EM Reconstruction of the CSM Complex**

(A) Raw micrograph, with representative single particles in white circles.

(B) Class averages and reprojections from the 3D reconstruction.

(C) Surface representation of the full 3D CSM volume.

(D) FSC plot.

was confirmed by MS. None of the Cas6 paralogs present in *S. solfataricus* copurified with the complex, suggesting that Cas6 is not stably associated.

## Sequence Analysis of RNA Copurifying with CSM

The RNA copurifying with the CSM complex was isolated, end labeled, and analyzed by denaturing gel electrophoresis (Figure 1C). The RNA, which was remarkably defined in size at around 50 nt, was cloned and deep sequenced on an Illumina platform. From the 5.77 million reads of 36 nt obtained after filtering, 5.45 million (94%) could be mapped to the six CRISPR loci present in the *S. solfataricus* P1 strain from which the complex was purified (Lillestøl et al., 2006), suggesting highly specific uptake of crRNA by the CSM complex. The six CRISPR loci in *S. solfataricus* are designated with the letters A–F and are characterized by two different types of repeat sequence, the A and B repeats being significantly different from those of C, D, E, and F (Lillestøl et al., 2006). CSM-derived crRNAs from the A and B loci made up 89% of the total matches, which together constitute 32% of the total spacers present on the genome. The D, E, and F loci were significantly underrepresented, constituting 11% of the matches, in sharp contrast to the fact that they constitute 68% of the spacers in the genome (Table S1). On the contrary, deep sequencing of the CMR complex crRNA revealed a bias toward the C and D loci (Zhang et al., 2012). These biases may reflect functional coupling of the CSM and CMR complexes with different Cas6 paralogs that have complementary specificity for the two CRISPR repeat families present in *S. solfataricus*.

Deep sequencing revealed that, as observed previously for the crRNA component of the CMR complex (Zhang et al., 2012), crRNA begins with the repeat-derived 8 nt 5′ handle (Figure 1D). Spacers in *S. solfataricus* are quite variable in length, ranging from 34 to 48 nt with a median value around 39 nt (Lintner et al., 2011). Thus, in CSM, the "average" spacer of 39 nt will be bounded by 8 nt of repeat-derived 5′ handle and around 3 nt of repeat-derived 3′ handle (Figure 1D). The secondary cleavage of crRNA in this case may occur after binding to CSM, with the complex defining the final length of the crRNA. As observed previously for the crRNA from the CMR complex, there is considerable variation in the coverage of individual spacers in the sequencing data. For example, in locus C, spacers 2, 11, 17, 21, 29, 30, and 33 are highly represented whereas other spacers are represented at much lower levels (Figure 1E). There is no general trend toward higher coverage at the 5′ end of the array, which might be explained by higher levels of transcription of spacers nearer the promoter, as has been observed for *Pyrococcus furiosus* (Hale et al., 2012). The reasons for the variability observed may be a combination of differences in expression due to the presence of internal promoters in captured spacers, differences in the efficiency of processing by Cas6 due to spacer sequence or structure effects, or variability in the cloning efficiency.

## Electron Microscopy

To gain insights into the assembly of the CSM complex, we performed EM coupled to single-particle analysis. Individual images of the complex showed an elongated shape. Image classification allowed a first appreciation of a coiled structure, where two filaments are intertwined. Most particles fell on the EM grids on the long axis, in side or tilted views. Top views were, however, not included in the reconstruction because they might have been poorly stained as a result of the overall length of the complex. Three-dimensional (3D) reconstruction and analysis of CSM confirmed these initial observations, revealing an assembly formed by two intertwined protein filaments, one thicker than the other, connected by a wider base (Figure 2). The overall dimensions of the complex are 205×125×100 Å. The resolution of the final reconstruction was determined as ~30 Å, calculated by Fourier shell correlation with a 0.5 cutoff.

## Subunit Composition Probed by MS

In order to investigate the composition of the CSM complex, we carried out MS analysis. The complex purified with a 10× His-tag attached to the C terminus of the subunit Sso1428 or Sso1431 was first analyzed by denaturing high-performance liquid chromatography–mass spectrometry (HPLC-MS), which confirmed the presence of all eight subunits (Table S2). The RNA component was characterized by phenol extraction of the CSM

complex followed by ethanol precipitation (Hernández et al., 2009). An MS spectrum showed a single charge-state series with a mass measured as 16,520 Da, consistent with the 50 nt crRNA (assuming an average mass of 321.5 Da for the four major ribonucleotide residues). The unusual broadness of the charge-state peaks (Figure S1) most likely reflects the sequence heterogeneity of the crRNA. In addition, proteomics experiments identified a series of posttranslational modifications (PTM) in CSM subunits (Table S2). The most prominent PTM was methylation, present in all eight subunits. Extensive methylation of lysine residues in crenarchaea has been reported previously and is suggested to be an adaptation conferring enhanced protein thermostability (Botting et al., 2010). The small subunit (Sso1424) was found to be 15 amino acid residues shorter than the annotated sequence, beginning with an acetylated N-terminal Ser-16 and including a total of seven methylated lysines. Subunits Sso1425 and Sso1431 were also found to be phosphorylated. Recently, over 500 phosphoproteins from *S. solfataricus* have been identified, although the role of phosphorylation in this organism is not well understood (Esser et al., 2012). The measured masses of the Sso1426 and Sso1427 subunits were within 70 Da of one another (Table S2), precluding the possibility of discriminating between them in the MS experiments.

With the masses of the protein and RNA components established experimentally, we then recorded a MS spectrum for the intact complex. MS spectra for CSM preparations with a His-tag attached to either Sso1428 or Sso1431 were recorded under nondenaturing conditions. Spectra for both preparations were very similar, dominated by a single, well-resolved charge-state series at around 8,500 m/*z* (Figure 3A). The masses of the intact complexes tagged on Sso1431 and Sso1428 were measured as 427.7 and 427.6 kDa, respectively (Figure S2), indicating a stoichiometric existence for these subunits in the complex. Under the conditions employed, some dimers (855 kDa) of low intensity were observed, presumably due to the multiple occupancy of the complex within the final offspring droplets, which is an artifact of the electrospray process (Lane et al., 2009). Gas-phase dissociation of Sso1424, Sso1428, and Sso1426/7 was observed upon tandem MS (Figure 3B). These data suggest that the CSM complex exists as a homogeneous population comprising one single crRNA and eight distinct protein subunits, of which Sso1428 and Sso1431 are present in equimolar quantities.

The measured mass for the intact complex was 122 kDa higher than the sum of the masses of its constituent subunits and crRNA, suggesting that some subunits of CSM existed in multiple copies. To determine the subunit stoichiometry, we turned to quantitative proteomics, using a labeling approach. We selected representative tryptic peptides from each subunit for isotopic labeling at C-terminal R/K residues, and to ensure a 1:1 molar ratio the peptide from the largest subunit, we conjugated Sso1428 with the remaining peptides, resulting in eight dipeptides for synthesis (two for the subunit Sso1430; Table S3). Each synthetic dipeptide was individually spiked into the CSM preparation before trypsin digestion, and the resultant peptide mixtures were analyzed by liquid chromatography–mass spectrometry (LC-MS). Comparison of signals generated by the labeled peptides resulted in a list of ratios of Sso1428 relative to the other seven CSM subunits; uncertainties still existed, however, especially for the subunits Sso1424, Sso1425, and Sso1426 (Table 1). We therefore resorted to MS of the intact complex and performed an exhaustive mass search based on the intact mass measurement (427,611 Da). For this, we allowed flexibility of copy numbers of these three subunits by one, with the stoichiometry of remaining subunits fixed according to Table 1. The search resulted in only one hit within a mass error of 3% and thus unambiguously assigned the relative molar ratios of the eight CSM subunits Sso1424 to Sso1432 to be 3:1:4:1:1:1:1:1, with Sso1424 and Sso1426 present in three and four copies, respectively, and unit stoichiometry for the others (Table 1 and Table S4).

Having established its subunit composition and stoichiometry, we proceeded to investigate the organization of subunits within the intact complex. For this, we employed a combination of crosslinking (CXMS) and in-solution disassembly. The intact complex could be disrupted by decreasing the pH, and a series of subcomplexes sized from 357 kDa down to 120 kDa (Figures 3D–3F, species i–v) were formed. We employed tandem MS to assign the subcomplexes, revealing their compositions, all of which contained the largest subunit Sso1428 (Figure S3, Table S5). This allowed us to distinguish a stable "base" subcomplex comprising single copies of Sso1428, 1430, and 1431 and two copies of 1426 and 1427. Further dissociation of this subcomplex led to the hetero-dimer Sso1428:1430 (120 kDa).

This disassembly pattern allowed us to deduce an interaction map, with assistance from the characteristic EM structure, with an intertwined major and minor filament (Figure 2). Of the 13 CSM subunits, 12 form two filaments stemming from the large one, Sso1428 (Figure 3G). The minor filament (Sso1430–1425) contacts the base subunit via Sso1430 and dissociates first at acid pH. This was followed by loss of subunits Sso1432 and three copies of Sso1426, which constitute the bulk of the major filament. This loss correlated with the loss of the crRNA molecule, suggesting an important role for Sso1426 in crRNA binding. The order of the subunit interactions was further confirmed by chemical crosslinking with a Bis[sulfosuccinimidyl] suberate deuterated and nondeuterated pair to generate crosslinked peptides with a readily distinguishable isotopic signature. Over 100 crosslinks were identified, among which six repeatedly identified intersubunit links were considered (Table S6). These include the large subunit Csm1 (Sso1428) crosslinking with both Sso1430 and Csm4 (Sso1431), which supports the identification of these three subunits at the base of the CSM structure. At the head of the structure, the Sso1425 subunit crosslinked to both Sso1426 and Sso1432. A crosslink between Sso1424 and Sso1427 suggests that the two helical filaments contact one another near the base.

To explore the spatial arrangement of the subunits, we used ion mobility MS (IM-MS) to measure the collision cross sections (CCS) for the intact complex and subcomplexes (Figure 4A, Table 2). Experimental CCS values were used as restraints for structural characterization in which candidate models were scored by the closeness of fit between the experimental and calculated CCS values (Alber et al., 2005; Politis et al., 2010). A coarse-grained structural model for the CSM complex was generated this way, which is in good agreement with the EM map (Figures 4B–4D).

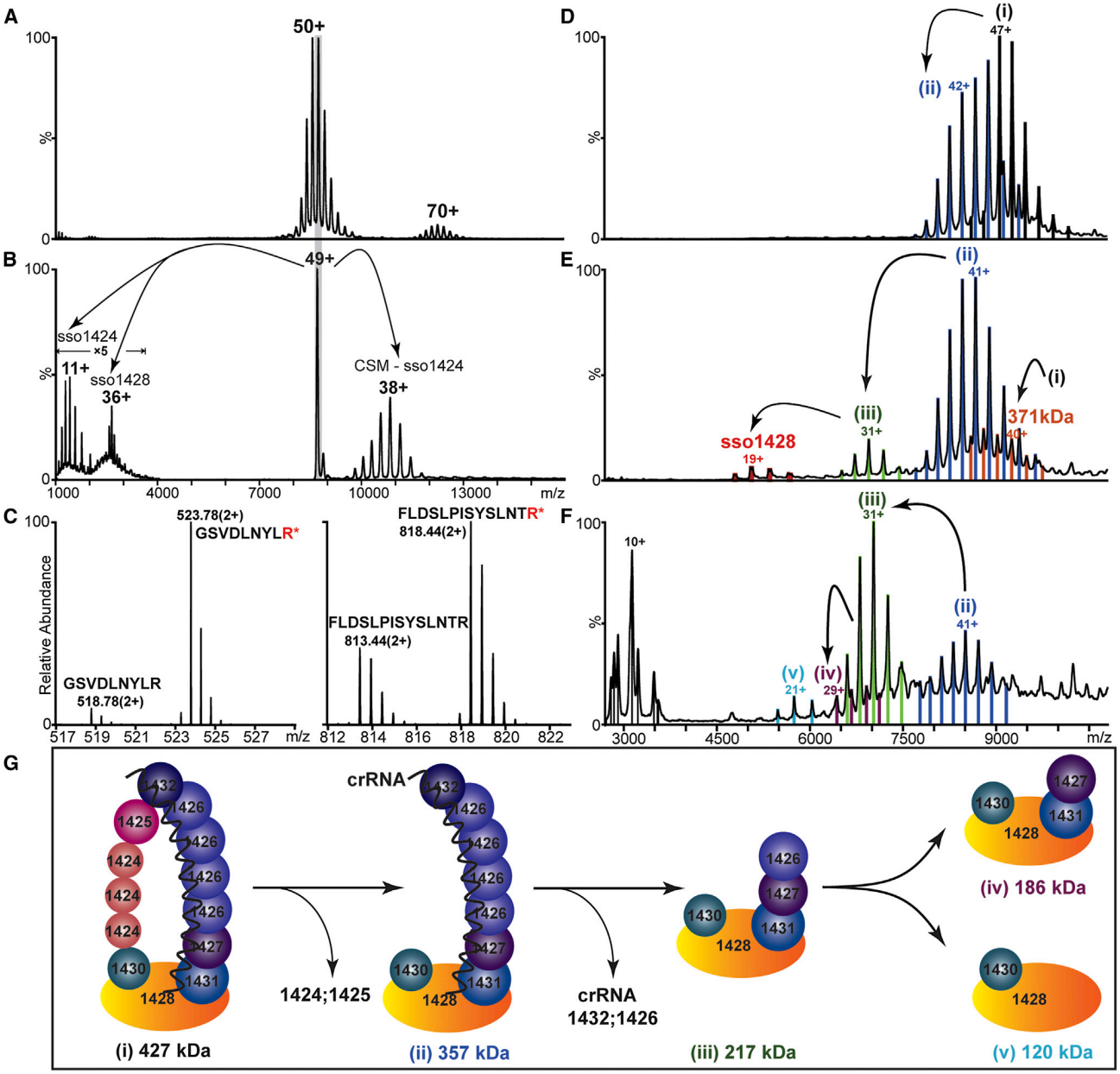



**Figure 3. MS Analysis of the CSM Complex Establishing Its Composition, Subunit Connectivity, and crRNA Binding**

(A) MS spectrum of the intact CSM reveals a well-resolved charge-state series at 8,500 m/*z* with a molecular mass of 427,789 Da, 122 kDa higher than the expected mass for a stoichiometric complex comprising eight subunits and one crRNA.

(B) The 49+ charge state of the complex was selected and subjected to acceleration, and dissociation of subunits Sso1424, Sso1428, and Sso1426/7 was observed by tandem MS.

(C) The molar ratio of Sso1426:Sso1428 was determined as 4:1 by relative quantification of tryptic peptides of Sso1426 and Sso1428 (GSVDLNYLR and FLDSLPISYSLNTR, respectively; see Table 1 and Table S3). Labeled peptides of the same sequences were synthesized and used as reference. ($^{15}$N,$^{13}$C)-labeled residues are colored red.

(D–F) Disassembly of the CSM complex resulted in a series of subcomplexes (i–v) in solutions of decreasing pH: 3.9 (D), 3.5 (E), and 3.2 (F).

(G) A complete CSM subunit interaction map was derived from MS data, including intact subcomplexes, crosslinking, and quantitative analysis (see also Figures S1–S3 and Tables S2–S6). The crRNA binds to subunits making up the major backbone and dissociates together with three copies of Sso1426 and Sso1432.

**Table 1. Quantification of CSM Subunits Relative to the Largest Subunit Sso1428**

| Subunits To Be Quantified | Selected Peptides | Ratio of Unknown Subunit:Sso1428 | | | | |
|---|---|---|---|---|---|---|
| | | Repeat 1 | Repeat 2 | Repeat 3 | Average | STD |
| Sso1432 | $^{18}$VGGGQEVGDNVIR$^{30}$ | 0.92 | 0.91 | 0.96 | 0.93 | 0.03 |
| Sso1431 | $^{293}$ISDLSSILNK$^{302}$ | 0.67 | 0.65 | 0.70 | 0.67 | 0.03 |
| Sso1430 | $^{150}$LLLYSILDLR$^{159}$ | 0.81 | 0.76 | 0.83 | 0.80 | 0.04 |
| Sso1430 | $^{199}$YLWEAENK$^{206}$ | 1.12 | 1.09 | 1.15 | 1.12 | 0.03 |
| Sso1426 | $^{136}$FLDSLPISYSLNTR$^{149}$ | 4.85 | 4.81 | 4.73 | 4.80 | 0.06 |
| Sso1425 | $^{62}$SLVESYTK$^{69}$ | 1.45 | 1.35 | 1.56 | 1.45 | 0.11 |
| Sso1427 | $^{129}$IFNPDPNR$^{136}$ | 0.80 | 0.79 | 0.83 | 0.81 | 0.02 |
| Sso1424 | $^{1}$N-acetyl-sSQDLLDIATR$^{11}$ | 3.62 | 3.51 | 4.03 | 3.72 | 0.27 |

### A Model for the CSM Complex Structure and Composition

The EM map of the CSM complex revealed an elongated structure, formed by two intertwined filaments connected at one end by a wide base (Figures 2 and 5). The level of detail obtained with 3D EM techniques allowed interpretation of the structure with fitting experiments. We built a backbone for the RAMP proteins on the basis of the Cas7 backbone present in the EMD-5314 map for the Cascade complex (Wiedenheft et al., 2011). Cas7 in Cascade is a larger polypeptide in comparison to the RAMP subunits present in CSM; therefore, we used only proximal domains, which are similar to RAMPs in size, to generate a backbone. We built a backbone using six Cas7 proximal domains (shown in light blue in Figure 5) that correspond to RAMP subunits Sso1427, 4 monomers of Sso1426, and Sso1432. At the base of the backbone, the Cas5 subunit from the bacterial Cascade complex (shown in dark blue in Figure 5), corresponding to Csm4 (Sso1431), is shown. This is consistent with volumetric observation, as well as with the CSM stoichiometry determined by MS. The pitch of the CSM backbone is identical to that of Cascade (Figures 5A–5D), whereas the CSM complex is slightly longer than Cascade (205 Å compared to 190 Å). The position of the RNA within this assembly remains elusive to EM at this resolution, but the thicker diameter of the major backbone is consistent with the presence of bound crRNA, and this corresponds to the binding orientation observed in Cascade. The thicker filament is ~130 Å long, in line with the size of the bound RNA. On both faces of the complex, the crevices between the two filaments (Figures 5A and 5C) have a width of ~24 Å and a length of ~130 Å. This is morphologically compatible with the diameter and length of a 38 bp DNA duplex (Figure S4), suggesting a possible role in target recognition at one of these two interfaces. This could also allow strand exchange with the crRNA bound along the Cas7 backbone. Consistent with this possibility, the purified CSM complex binds duplex DNA species with high affinity ($K_D$ around 100 nM), although sequence-specific binding could not be demonstrated because of the diversity of the crRNA bound to the complex (Figure S4). The size of the base of the structure is compatible with the expected volume of the full-length Cas10 (large) subunit. It should be noted that Cas10 could not fit within the density of the filaments, both of which are too thin to accommodate it. At the base of the helical backbone, the two structures are not comparable. This is consistent with the distinct structures of the large subunits of the type I and type III complexes, Cse1 and Cas10, respectively (reviewed in Reeks et al., 2013b).

## DISCUSSION

### Comparison with Other CRISPR Interference Complexes

Our data suggest that *S. solfataricus* CSM, and by extension all of the type IIIA complexes, are related structurally to type I complexes, sharing a crRNA-binding helical backbone built from Cas7-family RAMP domain proteins. In this case, the backbone interacts at one end with the Csm1-Csm4 (Cas10-Cas5) base domain, which may bind the 5′ end of the crRNA. This domain probably corresponds to the "crab claw" domain formed by the Cmr2 and Cmr3 subunits of the type IIIB complex (Zhang et al., 2012). Recent structures have shown that these two subunits form a deep crevice at their interface, which ends at the characteristic "cyclase" motif of the Cas10 subunit (Osawa et al., 2013; Shao et al., 2013). The structures reveal binding pockets for two nucleotides, which could represent part of a larger crRNA-binding site (Osawa et al., 2013). The conserved cyclase domain of Cas10 may thus play a role in recognition of the 5′ end of the crRNA rather than functioning as a catalytic domain. Additional biochemical studies are needed for investigation of this possibility.

The bulk of the crRNA-binding backbone is made up of four copies of Sso1426 and one of Sso1427, which can be regarded as Cas7 (or Csm3) family proteins. One end of the backbone is defined by an interaction at the base between the Cas7-like Sso1427 and the Cas5-like Sso1431, analogous to the Cas5-Cas7 core of type I complexes (Makarova et al., 2011b). The backbone is capped at the head by the Sso1432 and Sso1425 subunits, themselves RAMP family proteins, which presumably bind the 3′ end of the crRNA. Unlike the type IE complex, there is no 3′ crRNA hairpin structure and no integral Cas6 subunit. A second helical filament consisting primarily of three copies of the "small" subunit Sso1424 winds back down to link with the foot domain through the Sso1430 subunit. Recently, it has been suggested that the small subunits (Cse2, Cmr5, and Csm2) of all the type I and type III complexes are structurally related (Makarova et al., 2011a), and there are some structural data in support of this (Reeks et al., 2013a). However, there is no detectable sequence similarity between Sso1424 and the Csm2 subunits of CSM complexes from other

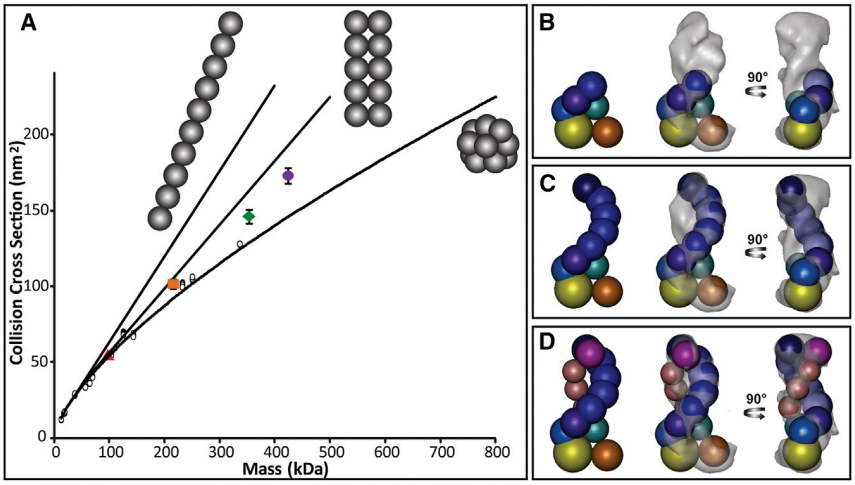



**Figure 4. Ion Mobility Measurement of the CSM Complex and Its Subcomplexes**

(A) CCS values measured for the intact complex (purple circle), the 357 kDa (green diamond) and 216 kDa (orange square) subcomplexes, and the largest subunit Sso1428 (red triangle) are plotted against their masses. Three trendlines are shown for linear, linear dimer, or collapsed "globular" conformations (left to right) for complexes composed of monomers (25 kDa). Considerable deviation from all conformation is evident for the intact complex and the two subcomplexes.

(B–D) Coarse-grain structural models, calculated for the intact complexes (D) and the 357 kDa (C) and 216 kDa (B) subcomplexes and fitted into the CSM EM map. Each subunit is represented by a sphere, sized proportionally to its mass, except that the largest Sso1428 is divided into two domains.

species such as *S. epidermidis*, let alone Cmr5 or Cse2 family proteins.

The similarity observed between the structures of the type I and type IIIA complexes is perhaps unsurprising given their similar function: both use bound crRNA to detect invading duplex DNA moieties, promoting strand exchange to form an R-loop that is a signal for DNA degradation. In contrast, the EM structure of the type IIIB (CMR) structure appears very different from that of the type IIIA complex, despite the fact that they share much clearer homology than either does with Cascade. The "body" of the CMR complex comprises a number of RAMP domain proteins (Cmr1, Cmr4, Cmr5, and Cmr6) that are assumed to bind RNA. However, they are not obviously arranged in the helical conformation seen for the type I and type IIIA complexes, instead appearing to form a more compact structure (Zhang et al., 2012). This may reflect the fact that CMR targets RNA substrates, which will not have the rigid helical structure of dsDNA. It remains to be seen whether all CMR complexes adopt this compact organization or whether this is specific to the crenarchaeal system.

## crRNA Binding and Processing in Type III Complexes

crRNA in *S. solfataricus* is generated by the cleavage of a primary pre-crRNA transcript within the repeat sequence by the Cas6 endonuclease (Reeks et al., 2013c; Shao and Li, 2013). This generates crRNA with a defined 8 nt repeat-derived 5′ handle, followed by a spacer sequence that can vary from 34 to 44 nt in length (Lintner et al., 2011) and a 3′ repeat-derived handle of 15–16 nt. This primary product is loaded, apparently without further processing, into the type IA complex (Lintner et al., 2011). However, in the type IIIB complex, further maturation was observed as generating shorter crRNAs with reduced 3′ ends (Zhang et al., 2012). In studies of the type IIIA system from *S. epidermidis*, mature crRNA of two sizes (39 and 45 nt) were observed. It has been proposed that crRNA is trimmed at the 3′ end by an unknown nuclease in a process directed by a ruler mechanism measured from the (Cas6-derived) 5′ end (Hatoum-Aslan et al., 2011).

Deep sequencing of the *S. solfataricus* CSM RNA complement confirmed that crRNAs were defined by a common 5′ end resulting from cleavage of the CRISPR repeat by Cas6, as expected. This suggests that, as observed previously for the *S. solfataricus* CMR and *S. epidermidis* CSM complexes, maturation involves 3′-end trimming. The most likely explanation may be that the complexes bind crRNA with an element of recognition of either the 5′ end or the 5′ handle sequence (or both), perhaps in the crevice formed by the Cas10 and Cas5 proteins as described above. Binding of crRNA by Cas7 family proteins results in the protection of a defined length of crRNA, and any excess is trimmed from the 3′ end by a nonspecific 3′-to-5′ exonuclease, as yet unidentified. In support of this, no mass shift was observed for the CSM complex treated with ribonuclease A, suggesting that the mature crRNA is fully protected by the complex (data not shown). The observation of two crRNA lengths differing by 6 nt in *S. epidermidis* CSM and *P. furiosus* CMR could be explained by differences in the number of Cas7-type crRNA-binding subunits present in the backbones of the complexes, as 6 nt approximates to the expected RNA-binding site size of Cas7 (Lintner et al., 2011). In other words, complexes with a 6-RAMP backbone would bind 36 nt of crRNA, while addition of a seventh RAMP subunit would allow the binding of a 42 nt crRNA. By contrast, *S. solfataricus* CSM appears to adopt a single, defined subunit composition with a single length of bound crRNA. It is possible that the control of backbone length by multimerization of RAMP proteins is not always precise.

## Target Degradation by Type IIIA Interference Complexes

The large (Cas10) subunits of the type IIIA and type IIIB complexes, Cmr2 and Csm1, each have an N-terminal HD-nuclease-like domain, reminiscent of that found in the Cas3 helicase-nuclease that is recruited for the degrading of viral DNA by Cascade. It was originally assumed that this would constitute the active site for all the type III complexes. However, this appears not to be the case for the *P. furiosus* CMR complex (Hale et al., 2012), and recent structural comparisons have highlighted the incomplete conservation of HD domains in all the type III complexes (Reeks et al., 2013b). Although CSM binds dsDNA

**Table 2. Collision Cross Sections of CSM Complex and Subcomplexes Measured by IM-MS**

| CSM (Sub-) Complexes | Mass (kDa) | Experimental CCS ($nm^2$) WH=32V WV=800$s^{-1}$ | WH=32V WV=700$s^{-1}$ | WH=30V WV=700$s^{-1}$ | Average | Calculated CCS (CG Model) | Difference (%) |
|---|---|---|---|---|---|---|---|
| Intact | 427 | 170.3 | 168.6 | 172.9 | 170.6 | 171.1 | +0.3 |
| Subcomplex I | 357 | 146.6 | 146.0 | 147.1 | 146.6 | 146.4 | −0.1 |
| Subcomplex II | 216 | 101.6 | 98.9 | 101.1 | 100.5 | 97.6 | −2.9 |
| Sso1428 | 97 | 55.1 | 55.0 | 56.5 | 55.6 | 56.1 | +0.9 |

CCS, collision cross sections.

with high affinity, we have so far been unable to demonstrate any crRNA-dependent nuclease activity for the type IIIA complex in vitro (C.R., J.Z., S.G., and M.F.W., unpublished data), and no other publication has reported such an activity, despite the fact that the complex was first reported to target DNA in vivo in 2008 (Marraffini and Sontheimer, 2008). One explanation is that, just as for Cascade, CSM is a surveillance complex that targets invading DNA and recruits a distinct nuclease to degrade targets. If so, the identity of this nuclease remains at present a matter for conjecture. Cas3 could in theory fulfill the role but is not always present in genomes harboring an active type IIIA system. The Csm6 protein is another possibility, although its structure bears more resemblance to families of transcription factors (Makarova et al., 2011b). It is conceivable that the nuclease varies in different lineages, which would be in keeping with the dynamic nature of the CRISPR system. Alternatively, the HD domain of the large subunit may be responsible for the degradation activity but be controlled in a manner that is not yet understood.

## Conclusions

This study has revealed clear similarities in the backbone structures of the CSM and Cascade surveillance complexes, suggesting a deep evolutionary relationship, as postulated from bioinformatics studies (Makarova et al., 2011a). Nonetheless, the differences should not be underestimated. For example, the requirement for a protospacer adjacent motif (PAM) in target sequences appears unique to the type I systems, and this may be reflected in the observation that the "large" subunits are not appreciably conserved between CSM and Cascade systems. Additional studies of the activity and mechanism of the CSM complex, both in vitro and in vivo, will be required in order to discern full details of role in the CRISPR system and its functional and structural relationship with Cascade.

## EXPERIMENTAL PROCEDURES

### Expression and Purification of Tagged CSM Complex in *S. solfataricus*

The gene encoding the large subunit of the complex, *sso1428*, was amplified with oligonucleotides containing *Nco*I and BamHI restriction sites. Ligation of the restricted PCR product into pMZ1 (Zolghadr et al., 2007) yielded plasmid pMZ-1428. Expression from pMZ1 leads to the addition of a C-terminal tandem tag (Strep and 10× His) to the protein. The expression cassette was excised from plasmid pMZ-1428 and ligated into the virus-based expression vector pSVA9, yielding plasmid pSVA-1428, which was transformed into the *S. solfataricus* PH1-16 expression strain, as described previously (Albers et al., 2006). After transformation, cells were first cultivated in unselective Brock medium containing 0.2% tryptone and 10 μg/ml uracil, then transferred to selective media containing 0.2% glucose and NZ-Amine without uracil. Once the $OD_{600nm}$ reached 0.6, cells were transferred to expression media containing 0.2% arabinose and NZ-amine to induce the expression of the tagged Sso1428 and then collected at an OD of 0.8–1.0. Later experiments involved the production of CSM complex tagged on subunit Sso1431 via the same methodology.

### Purification of Tagged CSM Complex from *S. solfataricus*

Cells were resuspended in buffer A (20 mM HEPES [pH 7.5], 250 mM NaCl, 30 mM imidazole) and disrupted by sonication for 6 × 3 min on ice. The lysate was centrifuged at 40,000 rpm for 45 min and loaded onto a Histrap column (GE Healthcare) equilibrated in buffer A. After being washed with 20 column volumes of buffer A, bound proteins were eluted with a linear gradient of buffer B (20 mM HEPES [pH 7.5], 250 mM NaCl, 1 M imidazole). Fractions containing the CSM complex were pooled, exchanged into buffer C (20 mM Tris·HCl [pH 8], 50 mM NaCl), and loaded onto a monoQ column (GE Healthcare) equilibrated with buffer D (20 mM Tris·HCl [pH 8], 50 mM NaCl, 1 mM EDTA, 1 mM DTT). Bound proteins were eluted with a linear gradient of buffer E (20 mM Tris·HCl [pH 8], 1 M NaCl, 1 mM EDTA, 1 mM DTT). Fractions containing the CSM complex were pooled, concentrated, and loaded onto a gel filtration column (S500, GE Healthcare) equilibrated with buffer F (20 mM Tris·HCl [pH 8], 150 mM NaCl). Fractions containing the CSM complex were pooled, concentrated, and stored at 4°C.

### Purification and Deep Sequencing of crRNA

RNA was extracted from the purified native CSM complex by the classical phenol/chloroform method followed by ethanol precipitation and vacuum desiccation. Dried RNA was resuspended in 5 μl of water and labeled in a 10 μl reaction containing polynucleotide kinase and 2 μCi $\gamma^{32}$P-ATP. Labeled RNAs were analyzed by electrophoresis on a 15% acrylamide, 7M urea, Tris-borate-EDTA (TBE) denaturing gel and visualized by phosphorimaging. Small RNA libraries were prepared with the use of the Small RNA Sample Prep Kit according to the manufacturers' instructions, starting from 100 ng RNA. The ligated RNA fragments were reverse transcribed, followed by ten cycles of PCR amplification. Subsequently, amplified libraries were purified on 6% polyacrylamide gels. The library was sequenced (36 bp single-read sequencing) with an Illumina Genome Analyzer IIx. Library preparation and sequencing was performed by the CNRS Imagif platform in Gif sur Yvette, France.This resulted in the addition of the adaptor sequence at the 3′ end of each sequence. Reads were processed, adaptor sequence was removed, and reads were mapped against the *S. solfataricus* P2 genome with the use of Galaxy (Blankenberg et al., 2010; Giardine et al., 2005; Goecks et al., 2010).

### Electron Microscopy

The CSM complex bound to crRNA was studied by negative-staining EM and single-particle analysis. Data were collected on an FEI F20 FEG microscope equipped with a 4k × 4k CCD camera. Images were collected under low-dose mode at a magnification of 29,000×, at a final sampling of 3.6 Å/pixel at the specimen level. Single-particle images were interactively selected with the Boxer program from the EMAN single-particle analysis package (Ludtke et al., 1999) and extracted into boxes. Image processing was performed with the IMAGIC-5 package (van Heel et al., 1996). The data set was resampled at 7.2 Å/pixel, and 7,829 images were band-pass filtered with a

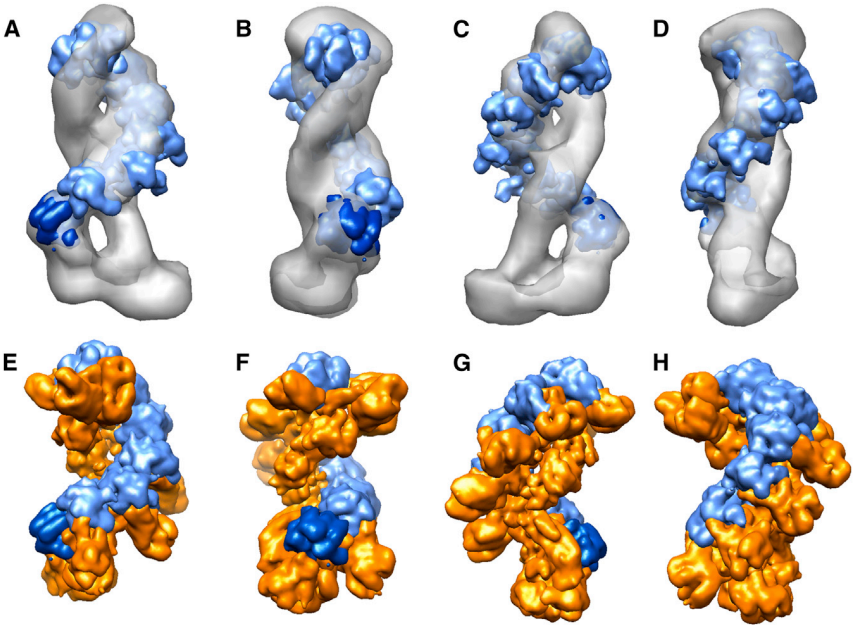



**Figure 5. Fitting the Cascade Backbone in CSM and Comparison of the Two Structures**

(A–D) Orthogonal views of CSM (gray surface) with fitted Cas5 (dark blue) and six Cas7 proximal domains (light blue).

(E–H) Orthogonal views of the Cascade complex from *E. coli*, where Cas5 and Cas7 proximal domains have been colored blue for direct comparison with CSM.

high pass cutoff of 110 Å and a low pass cutoff of 18 Å. The single-particle images were analyzed by multivariate statistical analysis with IMAGIC-5. The data set was subjected to successive rounds of alignment and classification in order to improve the resulting image class averages. We then generated a Gaussian blob, using the makeinitialmodel.py program from the EMAN package. The x, y, and z dimensions for the blob were chosen on the basis of the dimensions of class averages calculated with IMAGIC-5. Noise was added to the Gaussian blob with the use of the proc3d program in EMAN, to a 0.5 value. CSM class averages were aligned to the starting 3D volume by projection matching via the refine command in the EMAN package. The CMR/RNA structure was refined until the map converged. The resolution for the final reconstruction was calculated as ~30 Å through the use of the 0.5 FSC criterion. To interpret the map, we fitted a portion of the EMD-5314 map (Wiedenheft et al., 2011). To obtain the core Cas7 backbone, we segmented EMD-5314 using the Segger routine in Chimera and generated a volume containing six proximal Cas7 domains. The seventh module within the backbone was the Cas5 subunit. Figures were prepared with UCSF Chimera (Goddard et al., 2007).

## Mass Spectrometry

### *Electrospray Ionization LC-MS Analysis of CSM Subunits*

LC-MS analysis of individual CSM subunits was carried out on a Dionex Ultimate 3000 LC System (RSLCnano; Thermo) equipped with a 3 nl UV detector set at 214 and 280 nm. CSM was prepared in a 1:1 (v/v) mix of 0.1% TFA and 1 μl of sample applied to a PS-DVB reverse-phase monolithic column (Pepswift 100 μm i.d. × 25 cm; Thermo) equilibrated at 90% solvent A (0.05% TFA) and 10% solvent B (0.04% TFA, 90% ACN). A linear gradient of 10%–70% solvent B in 25 min at a flow rate of 600 nl/min was used. The column effluent was passed through a nanospray ionization interface into a QSTAR XL mass spectrometer (AB Sciex). For peptide analysis, the CSM complex was digested with tryspin (Promega). The resultant peptide mixture was separated on a reverse-phase C18 column (PepMap 75 μm i.d. × 50 cm; Thermo) before being analyzed on a LTO-Orbitrap XL hybrid mass spectrometer (Thermo). Eight proteins were identified as constituents of the CSM complex through a search against the NCBInr database using the Mascot search engine and are listed in Table S2.

### *Relative Quantification of CSM Subunits*

For quantification of the relative amount of each individual CSM subunits, the complete inventory of CSM tryptic peptides was surveyed. One or two peptides per subunit were selected for quantification according to previously published criteria (Schmidt et al., 2010). A library of synthetic dipeptides was then ordered from Thermo, containing each of these selected peptides combined with the sequence of a reference peptide (GSVDLNYLR) of subunit Sso1428. The dipeptides were isotopically labeled with ($^{15}$N;$^{13}$C) R/K residues to give a theoretical molar ratio of 1:1 and a mass increase (10/8 Da for R/K residues, respectively) for the component monopeptide upon trypsin cleavage. Subsequently, an aliquot of CSM complex was spiked with each of the synthetic dipeptide and the mixture was subjected to trypsin cleavage. The resulting digests were surveyed on the LTQ-Orbitrap. The extracted total ion chromatograms for the light and heavy peptide pairs were compared and their relative ratios calculated as quotients of the plotted peak areas.

### *Chemical Crosslinking of CSM Subunits Analyzed by MS*

The crosslinking experiment was initiated by mixing 2 μl of a 1:1 mixture of 12.5 mM deuterated (d4) and 12.5 mM nondeuterated (d0) BS3 crosslinkers with 20 μl aliquot of CSM complex at a concentration of 1 μg/μl. The reaction mixture was incubated for 1 hr at room temperature, and a control was prepared for comparison without addition of the crosslinkers. Potential crosslinked peptides were identified through the use of the MassMatrix Database Search Engine (Xu et al., 2008a; Xu et al., 2008b) and manually validated by (1) checking the presence of parent d4/d0 ion pairs in the MS spectra, (2) checking their absence in the control, and (3) checking qualities of the corresponding tandem MS spectra.

### *MS and IM-MS of the CSM Complex and Subcomplexes*

For MS of the intact complex, 20 μl of purified CSM (6 μg/μl) was exchanged into 200 mM AmAc buffer (pH 7.5) with the use of Micro Bio-Spin 6 Columns (Bio-Rad). The sample was diluted 1:10 into AmAc buffer, and 2 μl aliquots were electrosprayed from gold-coated borosilicate capillaries prepared in house. Spectra were recorded on a QSTAR XL (AB Sciex) modified for high mass detection (Sobott et al., 2002) and adjusted for the preservation of noncovalent interactions (Hernández and Robinson, 2007). MS experiments were performed at a capillary voltage of 1,200 V and declustering potentials of 40 V and 15 V. In tandem MS experiments, ions were isolated in the quadrupole and subjected to collision-induced dissociation (acceleration energy up to 200 V). For subcomplex generation, a 0.5 μl aliquot of the CSM solution was mixed with 19.5 μl of 200 mM AmAc containing incremental concentrations of acetic acid (5%–20% v/v) immediately before MS analysis.

All IM-MS spectra were recorded on a hybrid quadrupole (Q)-IM-ToF MS instrument known as Synapt G2 HDMS (Giles et al., 2011) and incorporating traveling-wave ion guide for IM separation (Waters). The instrument is modified for high mass transmission (Sobott et al., 2002) and uses nitrogen for mobility separation with the trap and transfer regions filled with argon. The Synapt G2 was operated at 3.21 mbar and 3.80 × $10^{-2}$ mbar for mobility and trap/transfer regions, respectively, which are separated by a "helium gate" pressurized at 1.41 bar. Ions were injected into the mobility cell at a 100 μs pulse with an injection voltage of 15 V. IM measurement for the CSM complex and subcomplexes was performed in triplicate, employing different combinations of wave height (WH) and wave velocity (WV) as follows: WH = 32V and WV = 800ms$^{-1}$; WH = 32V and WV = 700ms$^{-1}$; WH = 30V and WV = 700ms$^{-1}$.

### *Coarse-Grain Modeling of CSM*

An iterative series of modeling steps was employed for the CSM modeling combining information from MS and IM-MS, chemical crosslinking, and quantification experiments. First, each subunit (but the subunit of Sso1428 was divided into two domains) was represented as a sphere with a radius derived from its corresponding mass. We then employed a Monte Carlo sampling approach to build a large number of structures (10,000 models) for the CSM complex and subcomplexes consistent with the input connectivity data from MS-based experiments. Next, all generated models were scored and subsequently ranked on the basis of the violation of calculated CCSs values of model structures to the experimental values measured by IM. Finally, the top-scoring models were fitted into the EM map and the model with the best fit was selected as the final solution.

## ACCESSION NUMBERS

The raw sequence data have been uploaded to the Sequence Read Archive under accession number ERP003555. The CSM EM map has been submitted to the EMDB under accession number EMD-2420.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes six tables and four figures and can be found with this article online at http://dx.doi.org/10.1016/j.molcel.2013.08.020.

## ACKNOWLEDGMENTS

This work benefited from the facilities and expertise of the high-throughput sequencing platform of IMAGIF (Centre de Recherche de Gif). This work was funded by grants from the Biotechnology and Biological Sciences Research Council (BB/J005665/1 and BB/K000314/1 to M.F.W.; BB/J005673/1 to L.S. and M.F.W.), the Wellcome Trust (to M.Z. and C.V.R.), and the European Union 7th Framework Program PROSPECTS (Proteomics Specification in Space and Time) (HEALTH-F4-2008-201648 to A.P.), as well as an ERC Advanced Grant (to C.V.R.). The EM Facility at Edinburgh is supported by the Scottish Alliance for Life Sciences and the Wellcome Trust (WT087658MA). Giuseppe Cannone was the recipient of a Darwin Trust of Edinburgh Ph.D. studentship.

Received: June 4, 2013
Revised: July 11, 2013
Accepted: August 1, 2013
Published: October 10, 2013

## REFERENCES

Alber, F., Kim, M.F., and Sali, A. (2005). Structural characterization of assemblies from overall shape and subcomplex compositions. Structure *13*, 435–445.

Albers, S.V., Jonuscheit, M., Dinkelaker, S., Urich, T., Kletzin, A., Tampé, R., Driessen, A.J., and Schleper, C. (2006). Production of recombinant and tagged proteins in the hyperthermophilic archaeon Sulfolobus solfataricus. Appl. Environ. Microbiol. *72*, 102–111.

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. Curr. Protoc. Mol. Biol. *Chapter 19*, Unit 19.10.1–21.

Botting, C.H., Talbot, P., Paytubi, S., and White, M.F. (2010). Extensive lysine methylation in hyperthermophilic crenarchaea: potential implications for protein stability and recombinant enzymes. Archaea *2010*, 106341.

Esser, D., Pham, T.K., Reimann, J., Albers, S.V., Siebers, B., and Wright, P.C. (2012). Change of carbon source causes dramatic effects in the phosphoproteome of the archaeon Sulfolobus solfataricus. J. Proteome Res. *11*, 4823–4833.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., et al. (2005). Galaxy: a platform for interactive large-scale genome analysis. Genome Res. *15*, 1451–1455.

Giles, K., Williams, J.P., and Campuzano, I. (2011). Enhancements in travelling wave ion mobility resolution. Rapid Commun. Mass Spectrom. *25*, 1559–1566.

Goddard, T.D., Huang, C.C., and Ferrin, T.E. (2007). Visualizing density maps with UCSF Chimera. J. Struct. Biol. *157*, 281–287.

Goecks, J., Nekrutenko, A., and Taylor, J.; Galaxy Team. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. *11*, R86.

Hale, C.R., Majumdar, S., Elmore, J., Pfister, N., Compton, M., Olson, S., Resch, A.M., Glover, C.V., 3rd, Graveley, B.R., Terns, R.M., and Terns, M.P. (2012). Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. Mol. Cell *45*, 292–302.

Hatoum-Aslan, A., Maniv, I., and Marraffini, L.A. (2011). Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site. Proc. Natl. Acad. Sci. USA *108*, 21218–21222.

Hernández, H., and Robinson, C.V. (2007). Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. Nat. Protoc. *2*, 715–726.

Hernández, H., Makarova, O.V., Makarov, E.M., Morgner, N., Muto, Y., Krummel, D.P., and Robinson, C.V. (2009). Isoforms of U1-70k control subunit dynamics in the human spliceosomal U1 snRNP. PLoS ONE *4*, e7202.

Lane, L.A., Ruotolo, B.T., Robinson, C.V., Favrin, G., and Benesch, J.L.P. (2009). A Monte Carlo approach for assessing the specificity of protein oligomers observed in nano-electrospray mass spectra. Int. J. Mass Spectrom. *283*, 169–177.

Lillestøl, R.K., Redder, P., Garrett, R.A., and Brügger, K. (2006). A putative viral defence mechanism in archaeal cells. Archaea *2*, 59–72.

Lintner, N.G., Kerou, M., Brumfield, S.K., Graham, S., Liu, H., Naismith, J.H., Sdano, M., Peng, N., She, Q., Copié, V., et al. (2011). Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). J. Biol. Chem. *286*, 21643–21656.

Ludtke, S.J., Baldwin, P.R., and Chiu, W. (1999). EMAN: semiautomated software for high-resolution single-particle reconstructions. J. Struct. Biol. *128*, 82–97.

Makarova, K.S., Aravind, L., Wolf, Y.I., and Koonin, E.V. (2011a). Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. Biol. Direct *6*, 38.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J., Wolf, Y.I., Yakunin, A.F., et al. (2011b). Evolution and classification of the CRISPR-Cas systems. Nat. Rev. Microbiol. *9*, 467–477.

Marraffini, L.A., and Sontheimer, E.J. (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science *322*, 1843–1845.

Marraffini, L.A., and Sontheimer, E.J. (2010). Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature *463*, 568–571.

Osawa, T., Inanaga, H., and Numata, T. (2013). Crystal Structure of the Cmr2-Cmr3 Subcomplex in the CRISPR-Cas RNA Silencing Effector Complex. J. Mol. Biol., in press.

Politis, A., Park, A.Y., Hyung, S.J., Barsky, D., Ruotolo, B.T., and Robinson, C.V. (2010). Integrating ion mobility mass spectrometry with molecular modelling to determine the architecture of multiprotein complexes. PLoS ONE *5*, e12080.

Reeks, J., Graham, S., Anderson, L., Liu, H., White, M.F., and Naismith, J.H. (2013a). Structure of the archaeal Cascade subunit Csa5: relating the small subunits of CRISPR effector complexes. RNA Biol. *10*, 762–769.

Reeks, J., Naismith, J.H., and White, M.F. (2013b). CRISPR interference: a structural perspective. Biochem. J. *453*, 155–166.

Reeks, J., Sokolowski, R.D., Graham, S., Liu, H., Naismith, J.H., and White, M.F. (2013c). Structure of a dimeric crenarchaeal Cas6 enzyme with an atypical active site for CRISPR RNA processing. Biochem. J. *452*, 223–230.

Schmidt, C., Lenz, C., Grote, M., Lührmann, R., and Urlaub, H. (2010). Determination of protein stoichiometry within protein complexes using absolute quantification and multiple reaction monitoring. Anal. Chem. *82*, 2784–2796.

Shao, Y., and Li, H. (2013). Recognition and cleavage of a nonstructured CRISPR RNA by its processing endoribonuclease Cas6. Structure *21*, 385–393.

Shao, Y., Cocozaki, A.I., Ramia, N.F., Terns, R.M., Terns, M.P., and Li, H. (2013). Structure of the Cmr2-Cmr3 subcomplex of the Cmr RNA silencing complex. Structure *21*, 376–384.

She, Q., Singh, R.K., Confalonieri, F., Zivanovic, Y., Allard, G., Awayez, M.J., Chan-Weiher, C.C., Clausen, I.G., Curtis, B.A., De Moors, A., et al. (2001). The complete genome of the crenarchaeon Sulfolobus solfataricus P2. Proc. Natl. Acad. Sci. USA *98*, 7835–7840.

Sobott, F., Hernández, H., McCammon, M.G., Tito, M.A., and Robinson, C.V. (2002). A tandem mass spectrometer for improved transmission and analysis of large macromolecular assemblies. Anal. Chem. *74*, 1402–1407.

Söding, J., Biegert, A., and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. *33*(Web Server issue), W244-8.

van der Oost, J., Jore, M.M., Westra, E.R., Lundgren, M., and Brouns, S.J. (2009). CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem. Sci. *34*, 401–407.

van Duijn, E., Barbu, I.M., Barendregt, A., Jore, M.M., Wiedenheft, B., Lundgren, M., Westra, E.R., Brouns, S.J., Doudna, J.A., van der Oost, J., and Heck, A.J. (2012). Native tandem and ion mobility mass spectrometry highlight structural and modular similarities in clustered-regularly-interspaced shot-palindromic-repeats (CRISPR)-associated protein complexes from Escherichia coli and Pseudomonas aeruginosa. Mol. Cell. Proteomics *11*, 1430–1441.

van Heel, M., Harauz, G., Orlova, E.V., Schmidt, R., and Schatz, M. (1996). A new generation of the IMAGIC image processing system. J. Struct. Biol. *116*, 17–24.

Wiedenheft, B., Lander, G.C., Zhou, K., Jore, M.M., Brouns, S.J., van der Oost, J., Doudna, J.A., and Nogales, E. (2011). Structures of the RNA-guided surveillance complex from a bacterial immune system. Nature *477*, 486–489.

Xu, H., Yang, L., and Freitas, M.A. (2008a). A robust linear regression based algorithm for automated evaluation of peptide identifications from shotgun proteomics by use of reversed-phase liquid chromatography retention time. BMC Bioinformatics *9*, 347.

Xu, H., Zhang, L., and Freitas, M.A. (2008b). Identification and characterization of disulfide bonds in proteins and peptides from tandem MS data by use of the MassMatrix MS/MS search engine. J. Proteome Res. *7*, 138–144.

Yosef, I., Goren, M.G., and Qimron, U. (2012). Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Res. *40*, 5569–5576.

Zhang, J., Rouillon, C., Kerou, M., Reeks, J., Brugger, K., Graham, S., Reimann, J., Cannone, G., Liu, H., Albers, S.V., et al. (2012). Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. Mol. Cell *45*, 303–313.

Zolghadr, B., Weber, S., Szabó, Z., Driessen, A.J., and Albers, S.V. (2007). Identification of a system required for the functional surface localization of sugar binding proteins with class III signal peptides in Sulfolobus solfataricus. Mol. Microbiol. *64*, 795–806.

Structure
Article

# Crystal Structures of Malonyl-Coenzyme A Decarboxylase Provide Insights into Its Catalytic Mechanism and Disease-Causing Mutations

D. Sean Froese,[1,7] Farhad Forouhar,[2,7] Timothy H. Tran,[2,7] Melanie Vollmar,[1] Yi Seul Kim,[2] Scott Lew,[2] Helen Neely,[2] Jayaraman Seetharaman,[2] Yang Shen,[2] Rong Xiao,[3,4] Thomas B. Acton,[3,4] John K. Everett,[3,4] Giuseppe Cannone,[5] Sriharsha Puranik,[1] Pavel Savitsky,[1] Tobias Krojer,[1] Ewa S. Pilka,[1] Wasim Kiyani,[1] Wen Hwa Lee,[1] Brian D. Marsden,[1] Frank von Delft,[1] Charles K. Allerston,[1] Laura Spagnolo,[5] Opher Gileadi,[1] Gaetano T. Montelione,[3,4] Udo Oppermann,[1,6] Wyatt W. Yue,[1,*] and Liang Tong[2,*]

[1]Structural Genomics Consortium, University of Oxford, Oxford OX3 7DQ, UK
[2]Department of Biological Sciences, Northeast Structural Genomics Consortium, Columbia University, New York, NY 10027, USA
[3]Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, NJ 08854, USA
[4]Department of Biochemistry, Northeast Structural Genomics Consortium, Robert Wood Johnson Medical School, Piscataway, NJ 08854, USA
[5]Institute of Structural Molecular Biology, University of Edinburgh, Edinburgh EH9 3JR, UK
[6]NIHR Oxford Biomedical Research Unit, Botnar Research Centre, Oxford OX3 7LD, UK
[7]These authors contributed equally to this work
*Correspondence: wyatt.yue@sgc.ox.ac.uk (W.W.Y.), ltong@columbia.edu (L.T.)
http://dx.doi.org/10.1016/j.str.2013.05.001

## SUMMARY

**Malonyl-coenzyme A decarboxylase (MCD) is found from bacteria to humans, has important roles in regulating fatty acid metabolism and food intake, and is an attractive target for drug discovery. We report here four crystal structures of MCD from human, *Rhodopseudomonas palustris*, *Agrobacterium vitis*, and *Cupriavidus metallidurans* at up to 2.3 Å resolution. The MCD monomer contains an N-terminal helical domain involved in oligomerization and a C-terminal catalytic domain. The four structures exhibit substantial differences in the organization of the helical domains and, consequently, the oligomeric states and intersubunit interfaces. Unexpectedly, the MCD catalytic domain is structurally homologous to those of the GCN5-related *N*-acetyltransferase superfamily, especially the curacin A polyketide synthase catalytic module, with a conserved His-Ser/Thr dyad important for catalysis. Our structures, along with mutagenesis and kinetic studies, provide a molecular basis for understanding pathogenic mutations and catalysis, as well as a template for structure-based drug design.**

## INTRODUCTION

Malonyl-coenzyme A (malonyl-CoA) has long been established as the key intermediate in the biosynthesis of long-chain and very long-chain fatty acids (Wakil et al., 1983; Zammit, 1999), and it also has a crucial role in the regulation of fatty acid oxidation in mammals through its potent inhibition of carnitine palmitoyltransferase I (McGarry and Brown, 1997; Ramsay et al., 2001). Recent studies have demonstrated other important functions for this metabolite (Folmes and Lopaschuk, 2007; Lopaschuk et al., 2010; Saggerson, 2008), for example, in the regulation of food intake through its actions in the central nervous system (Fantino, 2011; Lane et al., 2008; Wolfgang and Lane, 2008) and in the control of fuel selection (carbohydrate versus fatty acids) in many tissues (Folmes and Lopaschuk, 2007; Saggerson, 2008). Therefore, malonyl-CoA may be a crucial regulator of energy homeostasis.

Cellular malonyl-CoA levels are controlled by several enzymes. Malonyl-CoA is produced by acetyl-CoA carboxylase (Cronan and Waldrop, 2002; Tong, 2013; Wakil et al., 1983) and is consumed by fatty acid synthase (Kuhajda, 2006), elongases (Guillou et al., 2010), and malonyl-CoA decarboxylase (MCD, E.C. 4.1.1.9) (Saggerson, 2008). The functional importance of malonyl-CoA suggests that modulators of these enzymes may have therapeutic applications. Hepatic overexpression of MCD in rats led to a decrease in circulating free fatty acid and, more importantly, alleviated insulin resistance normally induced by a high-fat diet (An et al., 2004). On the other hand, inhibition of MCD in the heart may be beneficial for treating cardiac ischemia and reperfusion (Ussher and Lopaschuk, 2009), which is supported by observations on MCD$^{-/-}$ mice (Dyck et al., 2006), as well as a collection of MCD inhibitors (Cheng et al., 2006a, 2006b, 2006c; Wallace et al., 2007). MCD inhibition has been found to be toxic to cancer cells, suggesting that it may be a target for anticancer therapy (Zhou et al., 2009). MCD inhibition can also reduce food intake and may be beneficial for obesity and diabetes treatment (Lopaschuk et al., 2010; Tang et al., 2010).

In mammals, MCD activity is found in the cytoplasm, mitochondria, and peroxisomes, and these different isoforms are encoded by a single gene (Courchesne-Smith et al., 1992; Gao et al., 1999; Joly et al., 2005; Sacksteder et al., 1999). MCD

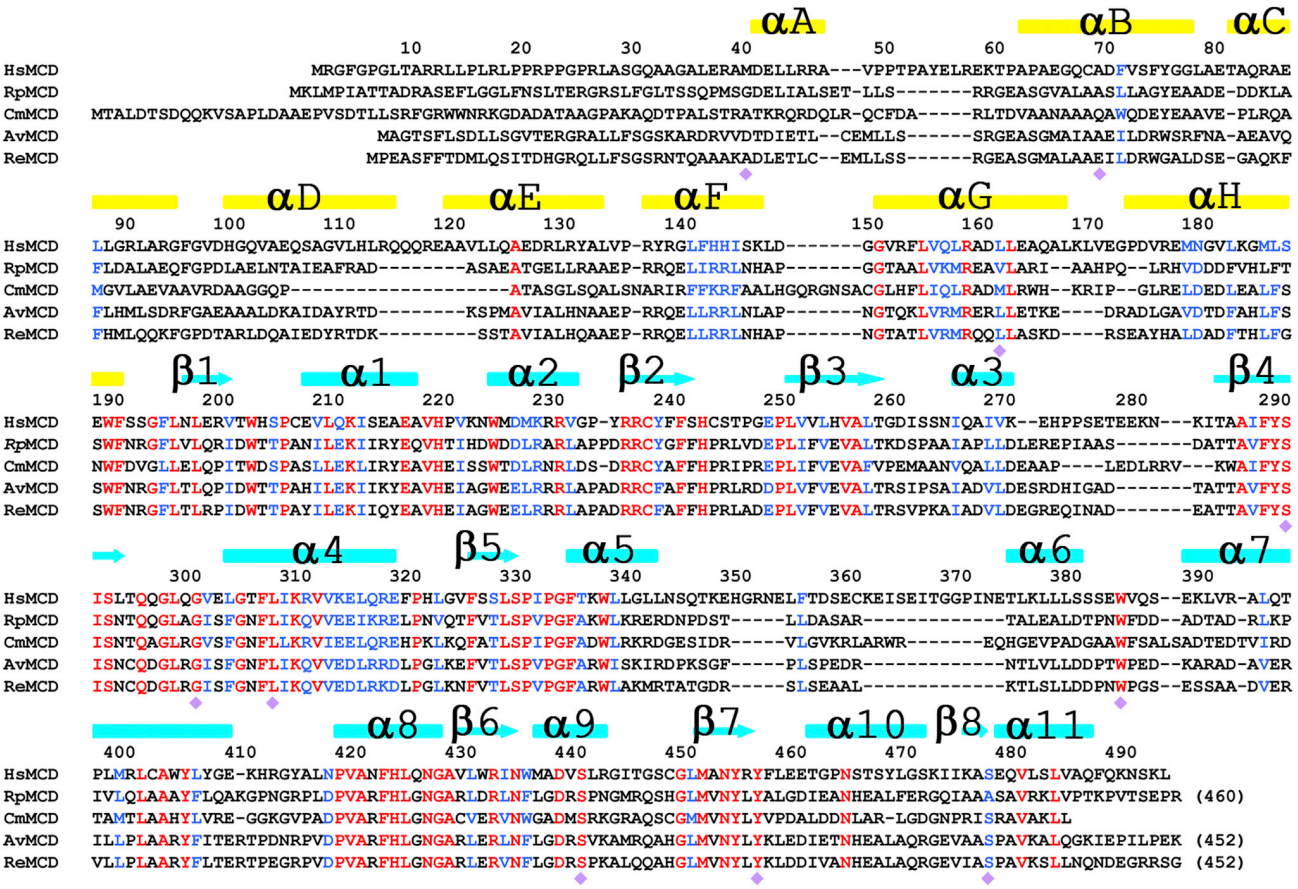



**Figure 1. Sequence Alignment of HsMCD, RpMCD, CmMCD, AvMCD, and ReMCD**

The secondary structure elements for HsMCD are indicated at the top of the alignment, colored in yellow for those in the helical domain and cyan for those in the catalytic domain. Strictly conserved residues among the five sequences are shown in red and highly conserved residues in blue. The purple diamonds indicate sites of disease-causing missense mutations in HsMCD.

deficiency in humans (Mendelian Inheritance in Man No. 248360), a rare autosomal recessive disorder, is characterized by malonic aciduria, developmental delay, cardiomyopathy, and neonatal death in severe cases (Malvagia et al., 2007; Salomons et al., 2007; Xue et al., 2012), supporting the important role of this enzyme in cellular functions. There is, as yet, no genotype-phenotype correlation for the ~30 pathogenic mutations identified (Xue et al., 2012).

MCD (~50 kDa) is also found in bacteria, plants, and other organisms with conserved amino acid sequences (Figure 1). For example, human MCD (HsMCD) and *Rhodopseudomonas palustris* MCD (RpMCD) share 34% sequence identity, while RpMCD and *Rhizobium etli* MCD (ReMCD) share 56% sequence identity (Figure 1). MCDs belong to the PFAM domain family PF05292 but do not share recognizable homology with other proteins in the sequence database, including methylmalonyl-CoA decarboxylase (Benning et al., 2000) and other decarboxylases. Purification of several animal and bacterial MCDs have been reported over the years (Kim and Kolattukudy, 1978; Kolattukudy et al., 1981; Lee et al., 2002; Lo et al., 2008; Zhou et al., 2004), and the crystallization of a bacterial MCD was also reported (Jung et al., 2003). However, no crystal structure was available on any of the MCDs, and the catalytic mechanism is still poorly understood.

We report here the crystal structures of human MCD as well as three bacterial MCDs at up to 2.3 Å resolution. The MCD monomer contains an N-terminal helical domain and a C-terminal catalytic domain, and the catalytic domain shares unexpected structural homology to the GCN5-related *N*-acetyltransferase (GNAT) superfamily. The N-terminal helical domain is involved in the oligomerization of MCDs, although there are substantial differences in the organization of the dimers and tetramers among MCD orthologs.

## RESULTS AND DISCUSSION

### Structure Determination

Wild-type HsMCD (residues 40–491, corresponding to the mature mitochondrial form) failed to crystallize. Adopting the surface entropy reduction (SER) strategy (Cooper et al., 2007), two charged patches in HsMCD, Glu58-Lys59 and Glu278-Glu279-Lys280, were predicted to be surface-exposed by the SER prediction server (http://services.mbi.ucla.edu/SER/; Goldschmidt et al., 2007), and site-directed mutagenesis was used to

**Table 1. Summary of Crystallographic Information**

| Structure | HsMCD | RpMCD | AvMCD | CmMCD |
|---|---|---|---|---|
| Space group | $C222_1$ | $P2_12_12$ | $I4_122$ | $C2$ |
| Unit cell parameters ($a$, $b$, $c$, $\alpha$, $\beta$, $\gamma$) | 95.6, 175.3, 151.8, 90, 90, 90 | 141.5, 159.8, 108.6, 90, 90, 90 | 100.4, 100.4, 242.7, 90, 90, 90 | 191.0, 69.4, 74.4, 90, 103.8, 90 |
| Resolution range for refinement (Å)[a] | 30–2.8 (2.9–2.8) | 30–2.7 (2.8–2.7) | 30–3.1 (3.2–3.1) | 30–2.3 (2.4–2.3) |
| Number of observations | 495,940 | 627,249 | 110,903 | 163,015 |
| $R_{merge}$ (%) | 12.5 (106.4) | 6.0 (61.2) | 10.5 (55.1) | 6.3 (44.1) |
| Redundancy | 5.0 (5.0) | 4.7 (4.4) | 5.2 (4.8) | 3.7 (3.5) |
| I/σI | 8.4 (1.6) | 25.2 (2.4) | 15.9 (2.6) | 23.0 (2.7) |
| Number of reflections | 31,694 | 123,627 | 19,052 | 37,613 |
| Completeness (%) | 100 (100) | 95 (85) | 89 (70) | 89 (72) |
| $R$ factor (%) | 21.2 (25.6) | 22.5 (34.0) | 22.0 (26.1) | 23.9 (28.6) |
| Free $R$ factor (%) | 25.5 (29.5) | 27.9 (38.3) | 29.1 (34.1) | 28.6 (33.3) |
| rms deviation in bond lengths (Å) | 0.010 | 0.007 | 0.009 | 0.007 |
| rms deviation in bond angles (°) | 1.1 | 1.3 | 1.4 | 1.2 |

[a]The numbers in parentheses are for the highest resolution shell.

**Table 2. Summary of Kinetic Parameters on Human MCD**

| Enzyme | $K_m$ (μM) | $k_{cat}$ ($s^{-1}$) | $k_{cat}/K_m$ ($M^{-1}s^{-1}$) |
|---|---|---|---|
| Wild-type HsMCD | 38 ± 12 | 33 ± 2 (1)[a] | $8.7 \times 10^5$ (1) |
| Quintuple SER mutant | 58 ± 17 | 45 ± 4 (0.73) | $7.8 \times 10^5$ (1.1) |
| H423N | 32 ± 4 | 4.7 ± 0.1 (7.0) | $1.4 \times 10^5$ (6.2) |
| S329A | 19 ± 4 | 0.30 ± 0.01 (110) | $1.5 \times 10^4$ (58) |
| Y456S | 132 ± 19 | 44 ± 2 (0.75) | $3.3 \times 10^5$ (2.6) |
| S290F | 37 ± 5 | 15 ± 1 (2.2) | $4.1 \times 10^5$ (2.1) |

[a]The ratio for values between the wild-type and mutant enzymes are given in the parentheses.

substitute alanine for each of these residues simultaneously. The structure of the E58A/K59A/E278A/E279A/K280A quintuple mutant was determined by single isomorphous replacement with anomalous scattering and refined at 2.8 Å resolution (Table 1; Figure S1 available online). The mutant exhibited similar oligomeric and enzymatic properties as wild-type HsMCD (Table 2). Inspection of the structure revealed both alanine-substituted patches to be located in surface-exposed regions: Glu58-Lys59 was found in the loop connecting helices αA and αB, while the loop containing residues 278–280, connecting strands β3 and β4, was disordered.

Bacterial MCDs were targeted as part of the broad program of the National Institutes of Health (NIH) Protein Structure Initiative on structural coverage of large protein domain families (Liu et al., 2007). We obtained crystals for several bacterial MCDs, but most of them showed poor diffraction quality (about 5 Å resolution). After significant efforts at optimization and diffraction screening, we collected X-ray diffraction data for RpMCD, *Agrobacterium vitis* MCD (AvMCD), and *Cupriavidus metallidurans* MCD (CmMCD) at up to 2.3 Å resolution. We solved the structure of RpMCD by the selenomethionyl single-wavelength anomalous diffraction method and the structures of AvMCD and CmMCD by molecular replacement (Table 1).

## Structures of MCD Monomers

The structures of the monomers of HsMCD (Figure 2A), RpMCD (Figure 2B), AvMCD (Figure 2C), and CmMCD (Figure 2D) can be divided into two domains: an N-terminal helical domain (130–150 residues) and a C-terminal catalytic domain (270–300 residues) connected via a short linker peptide. Consistent with this two-domain organization, the sequence conservation among the MCDs also appears to be bipartite (Figure 1). For example, the catalytic domains of HsMCD and RpMCD share 40% sequence identity, while their helical domains have only 24% identity. The N-terminal domain of HsMCD and several other MCDs are rich in Leu residues, which are concentrated in the helical segments.

The helical domain contains a bundle of six helices (αA–αC, αF–αH; Figures 2A–2D and S2). Helices αA and αB, and αG and αH form antiparallel hairpins and are arranged somewhat similar to those in armadillo/Huntington, elongation factor 3, protein phosphatase 2A, the yeast kinase TOR1 (HEAT), and tetratricopeptide repeats. However, the intervening helices αC and αF are located away from each other and run almost perpendicular to the other four helices. In addition, there is an insert of a helical hairpin (αD and αE) between helices αC and αF, which projects ~30 Å away from the rest of the monomer (Figure S1). This helical hairpin insert as well as the helical domain itself helps mediate the oligomerization of MCD (see below).

The catalytic domain of MCD contains a central eight-stranded, mostly antiparallel β sheet (β1–β8) that is surrounded by at least 11 α helices (α1–α11; Figures 2A–2D). Strands β4 and β5 in the middle of the β sheet, the only two neighboring strands that are parallel to each other (in a β-α-β motif), are splayed apart from each other at their C-terminal ends, and the active site of the enzyme is located in this region (see below). There is an insert of three additional helices (α5–α7) between strands β5 and β6 in HsMCD, RpMCD, and AvMCD, while CmMCD has an insert of five helices here. The sequences of this insert are poorly conserved among the MCDs (Figure 1).

The overall structures of the catalytic domains are similar, with root-mean-square (rms) distance of 1.2–1.5 Å for equivalent Cα atoms located within 3 Å of each other between any pair of the four structures. This structural similarity is particularly high for the central β sheet of the catalytic domain, as illustrated for overlays between HsMCD and RpMCD (Figure 2E), HsMCD and CmMCD (Figure 2F), and other structure pairs (Figure S1). On the other hand, many of the helices of the catalytic domain, especially those in the insert between β5 and β6, have large positional differences. Moreover, with the catalytic domains in overlay, significant differences in the orientation and position of the N-terminal helical domain are observed among the MCDs, corresponding to relative rotations of 15°–25° (Figures 2E, 2F, and S3). In addition, the helical hairpin insert between αC and αF is absent in CmMCD (Figures 2D and S2).

## Oligomeric Architectures of MCDs

HsMCD is a tetramer in solution based on gel filtration chromatography and analytical ultracentrifugation (AUC) studies

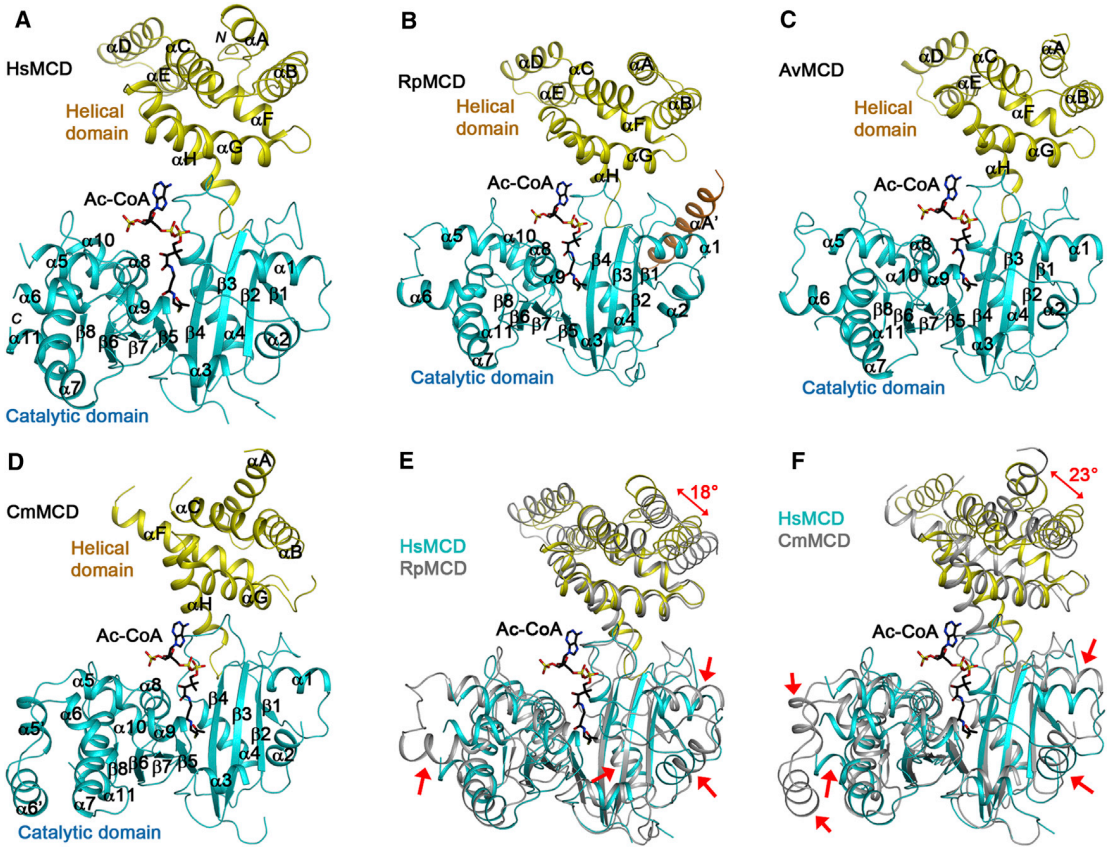



**Figure 2. Crystal Structures of MCD Monomer**

Schematic drawing of the structures of HsMCD (A), RpMCD (B), AvMCD (C), and CmMCD (D). The N-terminal helical domain is shown in yellow and the C-terminal catalytic domain in cyan. The bound position of acetyl-CoA in CurA (Gu et al., 2007) is shown as a stick model (in black). Overlays of the structures of HsMCD (in color) and RpMCD (in gray) (E) and HsMCD (in color) and CmMCD (in gray) (F). Regions of structural difference in the catalytic domain are highlighted with the red arrows. The difference in the orientations of the helical domains is also indicated. The structure figures were produced with PyMOL (http://www.pymol.org). See also Figure S1.

(Figure S2), consistent with the reported oligomerization state of many purified MCD enzymes. HsMCD sedimented in a single peak with an apparent molecular weight of ~200 kDa (Figure S2). The HsMCD crystal structure shows that the tetramer is made of a dimer of dimers (Figure 3A). A tight dimer of HsMCD is formed by extensive contacts of the helical domains of the two monomers, and the αD and αE helical inserts of the two monomers interact with each other in this dimer interface. Especially, helix αE of this insert contributes four leucine residues (122, 123, 129, and 133) to the interface. Approximately 1,800 Å$^2$ of the surface area of each monomer is buried in the dimer. Two HsMCD dimers then associate with each other through their catalytic domains, at ~60° angle for the planes of the two dimers (Figure S2), to form the tetramer with 222 symmetry. This interface primarily involves residues at the N-terminal end of the catalytic domain, burying ~500 Å$^2$ of the monomer surface area.

The architecture and shape of the HsMCD tetramer were also analyzed by electron microscopy coupled to single particle analysis. Images of negatively stained HsMCD contained a homogenous population of monodispersed single particles (Figure S2). Our three-dimensional (3D) reconstruction revealed a particle of 125 × 100 × 100 Å$^3$ in size with a central cavity, consistent in dimension and shape with the crystallographic tetramer (Figure 3B).

RpMCD and AvMCD are also tetramers in solution, based on multiangle static light scattering studies (data not shown). Like HsMCD, the RpMCD (Figure 3C) and AvMCD (Figure S2) tetramers are also dimer of dimers. However, the relative orientations of the dimers are substantially different (Figure S2). The central cavity of RpMCD tetramer also contains a helical segment (αA′) from the N terminus of two of the monomers (Figure 3C; Supplemental Information).

Surprisingly, CmMCD is a dimer in solution and the crystal structure reveals a completely different mode of dimerization as compared to HsMCD, RpMCD, and AvMCD. The two CmMCD monomers associate in a head-to-tail fashion such that the N-terminal helical domain of one monomer is in contact with the C-terminal catalytic domain of the other monomer, including the helical insert between strands β5 and β6 (Figure 3D). Approximately 1,100 Å$^2$ of the surface area of each monomer is buried in this dimer.

The variations in the oligomers of MCDs are likely due to the differences in the conformations of the N-terminal helical domains and the positions of these domains relative to the catalytic

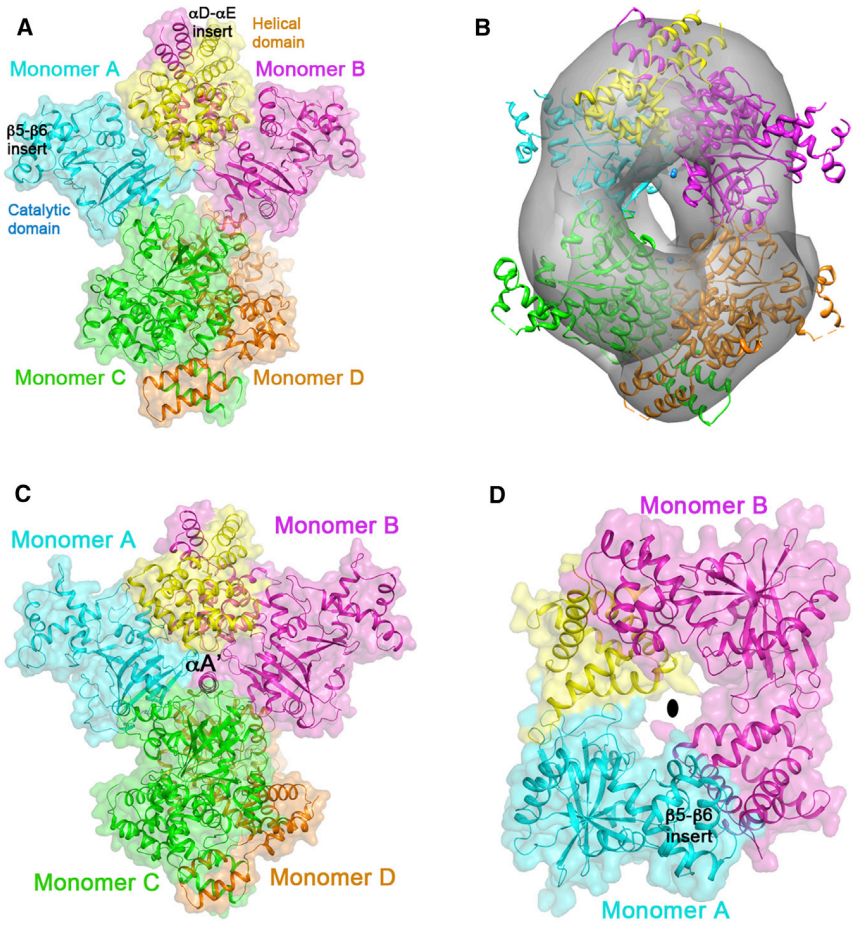



**Figure 3. The Oligomers of MCD**
(A) Structure of the HsMCD tetramer. A semi-transparent surface of the structure is also shown.
(B) Docking of the HsMCD tetramer structure into the EM reconstruction.
(C) Structure of the RpMCD tetramer.
(D) Structure of the CmMCD dimer. The 2-fold axis of the dimer is indicated with the oval (black).
See also Figure S2.

domains. For example, clear differences are visible between the HsMCD and RpMCD dimers (Figure S2), thereby affecting their tetramer formation. CmMCD lacks the helical insert in the helical domain and has two additional helices between β5 and β6 in the catalytic domain (Figure 2D), which may explain why it cannot form a similar dimer and tetramer as HsMCD or RpMCD.

While this paper was under review, a structure of HsMCD at 3.29 Å resolution was reported (Aparicio et al., 2013). The overall structures of the HsMCD monomers in the two reports are similar, with rms distance of 1.5 Å for 380 equivalent Cα atoms (Figure S2). There are recognizable differences in the organization of the dimer and tetramer between the two structures, although the overall architectures of the two tetramers are similar (Figure S2).

## Unexpected Structural Homology to GNAT Enzymes

The structure of the MCD catalytic domain unexpectedly shows strong homology to proteins belonging to the GNAT superfamily (Dyda et al., 2000; Neuwald and Landsman, 1997; Vetting et al., 2005), based on a Protein Data Bank (PDB) search with the program DaliLite (Holm et al., 2008). As the name indicates, most of these enzymes are *N*-acetyltransferases, a catalytic activity highly distinct from that of MCD. On the other hand, the overall backbone folds of these enzymes are homologous. GNAT proteins typically contain a seven-stranded β sheet, which corresponds to the first seven strands in the catalytic domain of MCD, with the splaying of the β4 and β5 strands a common feature among these structures. The sequence conservation between MCDs and these other GNAT members is, however, much lower, around 10% for structurally equivalent residues. As expected, the catalytic machinery in the active site is also distinct between MCD and the *N*-acetyltransferases.

The closest structural homolog, with a *Z* score of 16.6 from DaliLite, is the catalytic domain of the loading module of the polyketide synthase for curacin A (CurA) from *Lyngbya majuscula*, a GNAT protein that was shown not to have *N*-acetyltransferase activity (Gu et al., 2007; Figures 4A and 4B). Instead, this loading module harbors both malonyl-CoA decarboxylase and acetyl *S*-transferase activities. Despite the 13% identity for structurally equivalent residues between the two proteins, the catalytic residues for the decarboxylase activity of CurA are conserved in MCD (see below).

The N-terminal helical domain of MCDs does not have a counterpart in the GNAT enzymes. Consequently, the modes of oligomerization of MCDs are entirely different from these other GNAT enzymes. GNATs typically exist as monomers or dimerize via their GNAT core, and the predominant dimerization mode is by juxtaposing the GNAT β strands from both subunits to form a continuous β sheet. In contrast, the GNAT β strands in MCDs are not available for dimerization due to the presence of the large helical insert between strand β5 and β6. MCD dimerization is instead mediated by the N-terminal helical domain.

MCD represents a second example where a GNAT protein possesses a catalytic activity distinct from *N*-acetyltransferase. At the same time, the different activities of these GNAT proteins share the common substrate of acetyl- or malonyl-CoA. Therefore, the GNAT scaffold may have evolved to recognize the CoA moiety, and substitutions of several critical residues in the catalytic machinery may be sufficient to change the catalytic activity or substrate preference, such as succinyl-CoA (see below) (Vetting et al., 2008).

## The Active Site of MCD

Our extensive efforts to cocrystallize MCD with malonyl-CoA or acetyl-CoA have not been successful. Therefore, the structure

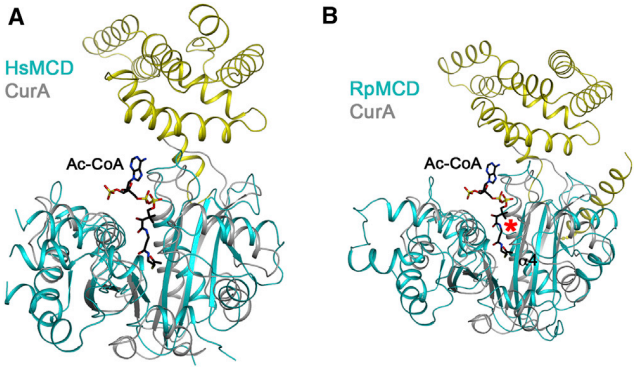



**Figure 4. Structural Conservation with CurA**

(A) Overlay of the structures of HsMCD (in color) and CurA (in gray). Acetyl-CoA in the CurA complex is shown as a stick model (black).

(B) Overlay of the structures of RpMCD (in color) and CurA (in gray). The red asterisk indicates large conformational differences in the N-terminal region of helix α4 between the two structures, which interacts with the phosphate groups of CoA in CurA.

of acetyl-CoA bound to CurA (Gu et al., 2007) was used as a guide for analyzing the MCD active site. This binding mode of acetyl-CoA is also generally similar to that in canonical GNAT enzymes, suggesting that the binding mode to MCD is likely to be similar as well.

The active site of MCD is located in a prominent groove in the surface of the monomer, where the most conserved residues among these enzymes are located (Figure 5A). The other monomers of the MCD oligomer make little, if any, contribution to the active site. For RpMCD, residues 55–58 in the other monomer of the dimer, in the loop linking the first two helices of the N-terminal domain, approach within ~10 Å of the expected position of the adenosine group in the active site. The equivalent loop in HsMCD is much longer, and Ala58 in this loop could have direct interactions with the adenine base of CoA. In the CmMCD dimer, the second monomer is located ~20 Å away from the active site.

The pantotheine group of CoA is positioned along strand β4 (Figures 5B and S3). The diphosphate and adenosine groups interact with the loop linking this strand to the following helix (α4) in HsMCD, and the diphosphate group also has favorable interactions with the dipole of this helix. In fact, this loop contains the signature sequence motif A in canonical GNAT enzymes (Neuwald and Landsman, 1997), (Q/R)xxGx(G/A)xxL, but the motif is not fully conserved in MCD, 299-(Q/R/A)xxxx(G/A)xxL-307 (Figure 1). Moreover, the loop and the following helix α4 are positioned differently in RpMCD (Figures 4 and S3) and CmMCD (Figure S1), suggesting that the binding mode of CoA to these MCDs may be somewhat different unless there is a conformational change upon CoA binding in these two enzymes. The 3′ phosphate group on the ribose of CoA is recognized by Arg387 in CurA (equivalent to Asn421 in HsMCD; Figure 5B). This residue is equivalent to Arg387 in RpMCD, which may have a similar function. However, this Arg residue is not conserved among the MCD enzymes. It shows variations to Asn in animal MCDs and His in some bacterial MCDs (Figure 1).

The acetyl group of acetyl-CoA interacts with conserved residues His389 and Thr355 in CurA (Figure 5B), which is proposed to be the catalytic dyad for its malonyl-CoA decarboxylase activity (Gu et al., 2007). The H389A, H389N, and T355V mutants have drastically reduced decarboxylase activity. The equivalent residues, His423 and Ser329 in HsMCD and His389 and Ser312 in RpMCD, are strictly conserved among the MCDs (Figure 1). In comparison, the His residue is equivalent to a Tyr residue in the canonical GNAT enzymes, which serves as the general acid for catalysis (Dyda et al., 2000; Neuwald and Landsman, 1997; Vetting et al., 2005). On the other hand, the Thr/Ser residue of CurA/MCD is not conserved in the canonical GNAT enzymes, while the general base for these enzymes, a Glu residue, is not conserved in CurA/MCD. These differences in the catalytic residues are likely the molecular basis for the distinct activity of CurA/MCD compared to the canonical GNATs.

The imidazole ring of His389 in CurA is held in place through a hydrogen bond with Tyr419. The equivalent residue in HsMCD, Tyr456, is also conserved among the MCDs. The carboxylate group of the malonyl-CoA substrate may lie over the surface of Phe288 in strand β4 (Figure 5B; HsMCD numbering), which is another strictly conserved residue among the MCDs (Figure 1).

The main chain of Thr355 in CurA has interactions with Arg404. However, this Arg residue is not conserved in RpMCD (Asp404), and in fact, an Asp residue is conserved at this position among the MCDs. The Arg404 residue may also be important for the acetyl *S*-transferase activity of CurA (Gu et al., 2007). The absence of this residue in MCD may be consistent with its lack of *S*-transferase activity.

Acetylation of Lys210, as well as mutation of Lys210 to Met, has been reported to inactivate rat MCD (Nam et al., 2006). Binding of acetyl-CoA protects rat MCD from the acetylation. In the HsMCD structure, the equivalent Lys211 side chain is on the surface of the tetramer, in a helix (α1) connecting strands β1 and β2, and ~20 Å from the active site. This side chain is mostly exposed to the solvent and does not have interactions with other conserved residues. Thus, it is not clear why this residue is essential for the catalysis by rat MCD.

To assess the functional importance of the active site His-Ser/Thr dyad of MCD, we carried out mutagenesis and kinetic studies with HsMCD. The S329A mutant of HsMCD had a 110-fold loss in $k_{cat}$ and 58-fold loss in $k_{cat}/K_m$, and the H423N mutant had a 7-fold loss in $k_{cat}$ (Table 2), consistent with their important roles in catalysis. In silico docking of malonyl-CoA into the HsMCD active site supports the kinetic data, showing that the substrate can position its thioester carbonyl (bridging the carboxylate leaving group and CoA backbone) in the vicinity (~3.2 Å) of Ser329 and His423 (Figure S3).

The reaction mechanism for MCD bears similarity to the acetyl transfer reaction of canonical GNATs, as they all need to polarize and stabilize the developing negative charge on the thioester carbonyl group (Figure 5C). Using HsMCD as example, we postulate that MCD proceeds through the formation of the tautomerized enolate intermediate, with the Ser329 and His423 dyad adopting important catalytic roles consistent with our docking and kinetic analysis. Phe288 may provide a nonpolar environment for the $CO_2$ leaving group, and the carbanion can abstract the proton from the side chain hydroxyl group of Ser329 acting as an acid (Figure 5C). This mechanism also has resemblance to that of a number of other CoA decarboxylases that do not employ cofactors (such as pyridoxal phosphate, thiamine, or

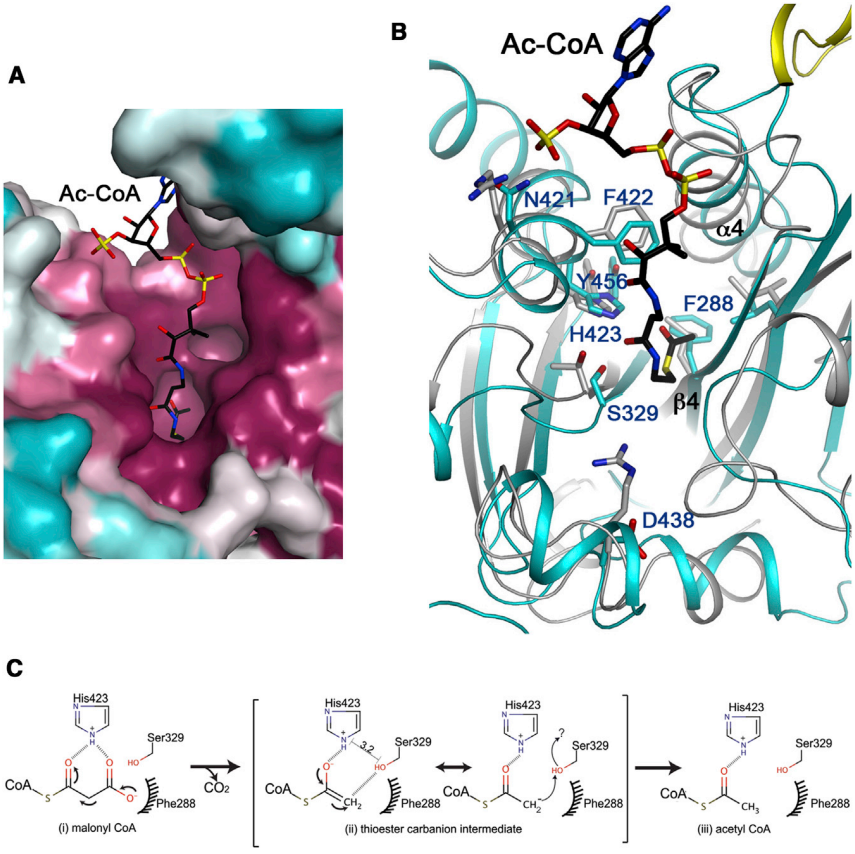



**Figure 5. The Active Site of MCD**

(A) Molecular surface of HsMCD in the active site region, colored by sequence conservation (magenta, most conserved; cyan, least conserved). The bound position of acetyl-CoA in CurA (Gu et al., 2007) is shown as a stick model (in black).

(B) An overlay of HsMCD (in color) and CurA (in gray) in the active site region. Side chains in HsMCD are labeled. The catalytic residues His423 and Ser329 of HsMCD are equivalent to His389 and Thr335 of CurA. Please see Figure S3 for a stereo version of this panel.

(C) Proposed catalytic mechanism for MCD (HsMCD numbering). Interatomic distance between His423 imidazole nitrogen and Ser329 hydroxyl oxygen is denoted in black line. Question mark represents possible proton transfer to reprotonate Ser329, from His423, a water molecule, or other unidentified sources.

See also Figure S3.

metal ions) to delocalize the buildup of the negative charge (Fu et al., 2004).

## Molecular Basis of Disease-Causing Mutations in MCD

The structure of HsMCD provides a molecular framework for understanding the impact of loss-of-function alleles in hereditary MCD deficiency. While the nonsense, frameshift, and deletion mutations result in truncated and thus nonfunctional proteins, the 11 known missense mutations (Table S1) are distributed throughout the structure with no discernible hot spot regions (Figure 6). The potential structural and biochemical consequences of these substitutions can be classified into three types. The first type is protein mistargeting and includes the two most N-terminal mutations, G3D and M40T, each of which lies within the predicted mitochondrial targeting sequence. Both mutations have been demonstrated to affect protein localization (Wightman et al., 2003). The second type of substitution likely disrupts protein folding through either protein instability or aggregation. These include A69V and L161P in the N-terminal helical domain, as well as W384C, S440I, and S477F in the catalytic domain. The third type involves substitutions in the GNAT core, affecting residues highly conserved among MCDs. These include S290F, G300V, L307R, and Y456S (Figure 6). Ser290 is located in strand β4 near the binding site for the CoA pantotheine moiety, though facing away from it. Mutation to the larger Phe residue would be expected to result in clashes with neighboring amino acids (His254 and Tyr289) and, hence, possible rearrangement of the active site and a partial loss of function. Indeed, the reconstituted S290F mutant showed a 2-fold decrease in $k_{cat}$ in vitro (Table 2). Gly300 and Leu307 are in the loop linking β4 and the following helix α4, being part of motif A. Both mutations result in substitution to larger residues that may clash with surrounding residues within this loop as well as residues on strand β3. Finally, Tyr456 interacts with the catalytic His423 residue (Figure 5C). Mutation to Ser would be expected to result in loss of His423 stabilization with consequent decreased substrate stability. In vitro, the Y456S mutant showed a 3.5-fold increased $K_m$ (Table 2), consistent with this proposal.

In summary, we report here structural information on MCD, revealing its catalytic machinery, oligomer organization, mechanism of disease-causing mutations, as well as unexpected homology to GNAT enzymes. The structural information should also facilitate the design and optimization of inhibitors against this enzyme. It has been suggested that the current inhibitors may require a hydrogen bond to a histidine residue for binding (Cheng et al., 2006c), and our structure suggests that this very likely is the catalytic His423 residue. Therefore, the active site of MCD is a promising target for the development of new therapeutic agents against human diseases.

## EXPERIMENTAL PROCEDURES

### Cloning, Expression, and Purification

A DNA fragment containing HsMCD (amino acids [aa] 40–491; IMAGE clone: 3357140) was subcloned into the pNIC28-Bsa4 vector (GenBank accession no. EF198106), incorporating an N-terminal tobacco etch virus (TEV)-cleavable $His_6$-tag. For surface entropy reduction, residues Glu58-Lys59 and Glu278-Glu279-Lys280 were replaced with Ala. The expression plasmids were transformed into *E. coli* BL21(DE3)-R3-pRARE2 cells, grown in Terrific broth medium with induction by 0.1 mM isopropyl-β-D-thiogalactopyranoside (IPTG) overnight at 18°C. Protein was purified by affinity (Ni-nitrilotriacetic acid; QIAGEN) and gel filtration (Superdex 200; GE Healthcare) chromatography.

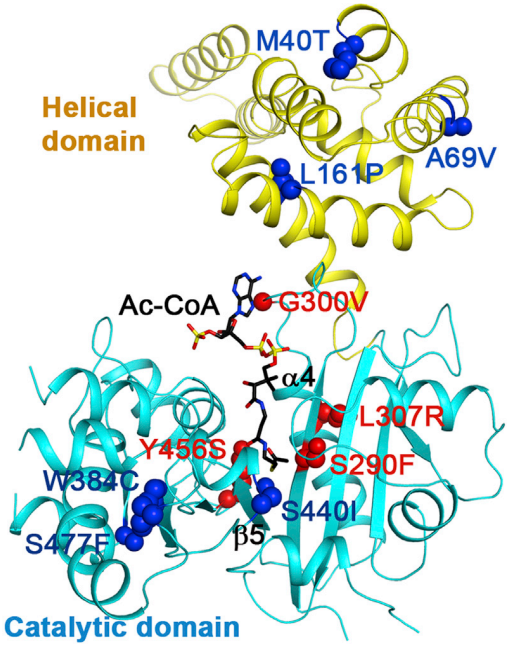



**Figure 6. Molecular Basis for MCD Disease-Causing Mutations**
The 11 missense pathogenic mutations (in red for those that could affect catalysis/substrate binding and blue for those that could affect folding/stability) are mapped onto the structure of HsMCD.
See also Table S1.

The production of the three bacterial MCDs, Rmet_2797 (CmMCD), RPA0560 (RpMCD), and Avi_5372 (AvMCD) from *Cupriavidus metallidurans*, *Rhodopseudomonas palustris*, and *Agrobacterium vitis*, respectively, was carried out as part of the high-throughput protein production process of the Northeast Structural Genomics Consortium (NESG) (Acton et al., 2005). The CmMCD, RpMCD, and AvMCD proteins correspond to NESG targets CrR76, RpR127, and RiR35, respectively. Full-length RpMCD and AvMCD were cloned into a pET21d (Novagen) derivative with C-terminal His-tag. Full-length CmMCD was cloned into pET26b with a C-terminal His-tag. *Escherichia coli* BL21 (DE3) pMGK cells, a rare codon enhanced strain, were transformed with each plasmid. A single isolate was transferred to 500 μl of Luria broth with ampicillin and kanamycin and incubated for 6 hr at 37°C. This preculture (40 μl) is then used to inoculate a 250 ml flask containing 40 ml of MJ9 minimal media (Jansson et al., 1996) and incubated overnight at 37°C. The entire volume of overnight culture is then used to inoculate a 2 l baffled flask containing 1.0 l of MJ9. The cultures are incubated at 37°C until the optical density at 600 nm reaches 0.8–1.0 units, equilibrated to 17°C, and induced with IPTG (1 mM final concentration) after addition of several amino acids to the medium to downregulate methionine synthesis (lysine, phenylalanine, and threonine at 100 mg/l; isoleucine, leucine, and valine at 50 mg/l; and L-selenomethionine at 60 mg/l) for 15 min (Doublié et al., 1996). In the case of CmMCD, the media contained methionine instead. Following overnight incubation, the cells were harvested by centrifugation. However, the full-length CmMCD, RpMCD, and AvMCD could not be purified this way, due to low expression and/or low solubility. Subsequently, construct optimization experiments revealed that expression of RpMCD, AvMCD, and CmMCD construct containing residues 8–451, 1–448, and 57–473, respectively, yielded soluble protein in each case without noticeable protein aggregation. The pET expression vectors for these constructs (NESG RpR127-8-451-21.13, NESG RiR35-1-448-21.13, and NESG ReR178-25-448-28) have been deposited in the Protein Structure Initiative Materials Repository (http://psimr.asu.edu).

Selenomethionyl RpMCD, AvMCD, and native CmMCD were purified by standard methods. Cell pellets were resuspended in lysis buffer (50 mM Tris [pH 7.5], 500 mM NaCl, 40 mM imidazole, and 1 mM Tris-(2-carboxyethyl) phosphine) and disrupted by sonication. The resulting lysate was clarified by centrifugation at 26,000 × *g* for 45 min at 4°C. The supernatant is then loaded onto an ÄKTAxpress system (GE Healthcare), and a two-step automated purification protocol is performed, comprised of a Ni-affinity column (HisTrap HP, 5 ml) and a gel filtration column (Superdex 75 26/60, GE Healthcare) in a linear series. A buffer containing 10 mM Tris (pH 7.5), 100 mM NaCl, 5 mM dithiothreitol (DTT), and 0.02% (w/v) $NaN_3$ is used for gel filtration. The purified Se-Met labeled RpMCD, AvMCD, and native CmMCD were concentrated to 11, 8, and 10 mg/ml, respectively, flash frozen in aliquots, and used for crystallization screening. Sample purity (>95%) and molecular weight were verified by SDS-PAGE and MALDI-TOF mass spectrometry, respectively.

### Protein Crystallization

Purified HsMCD (SER quintuple mutant) was concentrated to 10 mg/ml in a buffer containing 5 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) (pH 7.5), 100 mM NaCl, 1% (v/v) glycerol,l and 5 μM decanoyl-CoA. Crystals were obtained by sitting-drop vapor diffusion at room temperature by incubating protein in a 2:1 ratio with a precipitant containing 10% (w/v) polyethylene glycol (PEG) 20,000 and 0.1 M 2-(N-morpholino)ethanesulfonic acid (pH 6.0). The crystals belong to space group $C222_1$, with a dimer of HsMCD in the asymmetric unit. The tetramer is generated through a crystallographic 2-fold axis.

The purified Se-Met-labeled RpMCD, AvMCD, and native CmMCD were crystallized using microbatch method at 18°C. In the case of RpMCD and AvMCD, 2 μl of the protein solution containing 10 mM Tris (pH 7.5), 100 mM NaCl, 5 mM DTT, and 0.02% $NaN_3$ were mixed with 2 μl of the precipitant solution consisting of 0.1 M magnesium nitrate, 100 mM Tris (pH 8.5), and 33% (v/v) PEG 400 for RpMCD and 200 mM ammonium sulfate and 20% (w/v) PEG3350 for AvMCD. For CmMCD, 2 μl of the protein in a buffer consisting of 20 mM Tris (pH 7), 250 mM NaCl, 5% (v/v) glycerol, and 3 mM malonyl-CoA were mixed with a crystallization cocktail containing 160 mM magnesium chloride, 80 mM Tris (pH 8.5), 24% (w/v) PEG 4000, 20% (v/v) glycerol, and 3% (v/v) ethanol. The RpMCD and AvMCD crystals were cryoprotected by supplementing their respective crystallization cocktail with 20% (v/v) ethylene glycol and 20% (v/v) glycerol, respectively. No cryoprotecting solution was added into the crystallization cocktail containing CmMCD crystals for data collection at 100 K.

Crystals of RpMCD, AvMCD, and CmMCD belong to space group $P2_12_12$, $I4_122$ and $C2$, respectively, with four, one, and two molecules in the crystallographic asymmetric unit.

### Structure Determination and Refinement

For HsMCD, the structure was solved by multiple isomorphous replacement with anomalous scattering phasing. HsMCD crystals were derivatized with thimerosal or $K_2PtCl_4$ by 20 min incubation in reservoir solution supplemented with 5 mM of the respective heavy atom compound. X-ray diffraction data were collected at the Diamond Light Source beamlines IO2 and IO3 and processed and scaled with XDS (Kabsch, 2010) and Scala (Collaborative Computational Project, 1994), respectively. SHELXD (Sheldrick, 2008) identified three heavy atom sites in the mercury derivative. After including both derivatives in SHARP (Vonrhein et al., 2007) and subsequent density modification with SOLOMON (Abrahams and Leslie, 1996), substantial parts of the model were automatically built with BUCANEER (Cowtan, 2006). Manual model rebuilding was carried out with Coot (Emsley and Cowtan, 2004) and structure refinement with BUSTER (Global Phasing). No ligand electron density was observed in the active site. Residues 60–65, 115 and 116, 276–281, and 344–371, which represent surface-exposed regions in the structure, are disordered and not modeled.

The structure of RpMCD was determined by a single-wavelength anomalous diffraction data set to resolution 3.1 Å, which was collected at the peak absorption wavelength of selenium at the X6A beamline of the National Synchrotron Light Source. The diffraction images were processed with the HKL package (Otwinowski and Minor, 1997), and the selenium sites were located with the program SHELX (Sheldrick, 2008). SOLVE/RESOLVE was used for phasing the reflections and automated model building (Terwilliger, 2003). The majority of the model was built manually with the program XtalView (McRee, 1999). The structure refinement was performed with CNS 1.3 (Brünger et al., 1998).

The model thus obtained for RpMCD was used as a search model for structure determination of another data set of RpMCD to resolution 2.7 Å. The

model was subsequently used to determine structures of CmMCD and AvMCD to resolution 2.3 Å and 3.1 Å, respectively, using the molecular replacement method implemented in the program Molrep (Vagin and Teplyakov, 2000). The data processing and refinement statistics are summarized in Table 1.

### Decarboxylase Activity Measurement

MCD catalytic activity was determined following a published protocol (Kolattukudy et al., 1981). For HsMCD, the following reagents were added to a total of 100 μl in a 96 well plate: 50 mM HEPES (pH 7.5), 1 mM dithiothreitol, 5 mM L-malate, 1 mM nicotinamide adenine dinucleotide $(NAD)^+$, 0.1 mM reduced NAD, 1.925 U malate dehydrogenase (Sigma-Aldrich), 0.4 U citrate synthase (Sigma-Aldrich), 100-1000 nM HsMCD protein, and various concentrations (0 μM–500 μM) of malonyl-CoA. Absorbance changes at 340 nm were measured for 30 min and linear velocity used to calculate enzyme activity using GraphPad Prism (v.5.01).

### Analytical Ultracentrifugation

Sedimentation velocity (SV) experiments were performed in a Beckman Optima XL-I analytical ultracentrifuge (Beckman Instruments) using AnTi-50 rotor. Experiments were conducted at 30,000 rpm and 4°C using absorbance detection and cells loaded with 50 μM HsMCD in 10 mM HEPES (pH 7.5) and 150 mM NaCl. SV data were analyzed using SEDFIT (Schuck, 2000), while sedimentation coefficients, *s*, were calculated with SEDNTERP (Laue et al., 1992) version 1.09.

### Analytical Gel Filtration

Analytical gel filtration was performed on a Superdex 200 HiLoad 10/30 column (GE Healthcare) pre-equilibrated with 10 mM HEPES (pH 7.5) and 150 mM NaCl at a flow rate of 0.3 ml/min.

### Electron Microscopy

We studied the HsMCD assembly by negative staining electron microscopy and single particle analysis. Data were collected on a FEI F20 field emission gun microscope, equipped with an 8k × 8k charge-coupled device camera. Images were collected under low dose mode at a magnification of 50,000X at a final sampling of 1.6 Å/pixel at the specimen level. Single particle images were selected interactively using the Boxer program from the EMAN package (Ludtke et al., 1999). Image processing was performed using the IMAGIC-5 package (van Heel et al., 1996), and the single particle images were analyzed by multivariate statistical analysis. Selected class averages were used to calculate a starting 3D volume by common lines using the Euler program in the IMAGIC-5 package with no symmetry imposed. Manual fitting of the HsMCD tetramer was performed with UCSF Chimera (Goddard et al., 2007).

## ACCESSION NUMBERS

The PDB accession numbers for HsMCD, RpMCD, AvMCD, and CmMCD reported in this paper are 2YGW, 4KSA, 4KSF, and 4KS9, respectively.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Results and Discussion, three figures, and one table and can be found with this article online at http://dx.doi.org/10.1016/j.str.2013.05.001.

## ACKNOWLEDGMENTS

We thank Angela Lauricella and George DeTitta for setting up initial crystal screenings for the bacterial MCDs, Randy Abramowitz and John Schwanof for setting up the X4A beamline, Jean Jakoncic for setting up the X6A beamline, and the staff at the Diamond Light Source for help in synchrotron data collection. We also thank J. Elkins, S. Knapp, and P. Filippakopoulos for assistance in AUC experiments. The Structural Genomics Consortium is a registered charity (No. 1097737) funded by the Canadian Institutes for Health Research, the Canadian Foundation for Innovation, Genome Canada through the Ontario Genomics Institute, GlaxoSmithKline, Karolinska Institutet, the Knut and Alice Wallenberg Foundation, the Ontario Innovation Trust, the Ontario Ministry for Research and Innovation, Merck, the Novartis Research Foundation, the Swedish Agency for Innovation Systems, the Swedish Foundation for Strategic Research, and the Wellcome Trust. The Northeast Structural Genomics Consortium is funded by the Protein Structure Initiative of the NIH (U54 GM094597 to G.T.M. and L.T.). This research is also supported in part by a grant from the NIH (R01 DK067238 to L.T.).

Received: February 6, 2013
Revised: May 9, 2013
Accepted: May 9, 2013
Published: June 20, 2013

## REFERENCES

Abrahams, J.P., and Leslie, A.G.W. (1996). Methods used in the structure determination of bovine mitochondrial F1 ATPase. Acta Crystallogr. D Biol. Crystallogr. *52*, 30–42.

Acton, T.B., Gunsalus, K.C., Xiao, R., Ma, L.C., Aramini, J., Baran, M.C., Chiang, Y.W., Climent, T., Cooper, B., Denissova, N.G., et al. (2005). Robotic cloning and protein production platform of the Northeast Structural Genomics Consortium. Methods Enzymol. *394*, 210–243.

An, J., Muoio, D.M., Shiota, M., Fujimoto, Y., Cline, G.W., Shulman, G.I., Koves, T.R., Stevens, R., Millington, D., and Newgard, C.B. (2004). Hepatic expression of malonyl-CoA decarboxylase reverses muscle, liver and whole-animal insulin resistance. Nat. Med. *10*, 268–274.

Aparicio, D., Pérez-Luque, R., Carpena, X., Díaz, M., Ferrer, J.C., Loewen, P.C., and Fita, I. (2013). Structural asymmetry and disulfide bridges among subunits modulate the activity of human malonyl-CoA decarboxylase. J. Biol. Chem. *288*, 11907–11919.

Benning, M.M., Haller, T., Gerlt, J.A., and Holden, H.M. (2000). New reactions in the crotonase superfamily: structure of methylmalonyl CoA decarboxylase from Escherichia coli. Biochemistry *39*, 4630–4639.

Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr. D Biol. Crystallogr. *54*, 905–921.

Cheng, J.-F., Chen, M., Wallace, D., Tith, S., Haramura, M., Liu, B., Mak, C.C., Arrhenius, T., Reily, S., Brown, S., et al. (2006a). Synthesis and structure-activity relationship of small-molecule malonyl coenzyme A decarboxylase inhibitors. J. Med. Chem. *49*, 1517–1525.

Cheng, J.-F., Huang, Y., Penuliar, R., Nishimoto, M., Liu, L., Arrhenius, T., Yang, G., O'leary, E., Barbosa, M., Barr, R., et al. (2006b). Discovery of potent and orally available malonyl-CoA decarboxylase inhibitors as cardioprotective agents. J. Med. Chem. *49*, 4055–4058.

Cheng, J.-F., Mak, C.C., Huang, Y., Penuliar, R., Nishimoto, M., Zhang, L., Chen, M., Wallace, D., Arrhenius, T., Chu, D., et al. (2006c). Heteroaryl substituted bis-trifluoromethyl carbinols as malonyl-CoA decarboxylase inhibitors. Bioorg. Med. Chem. Lett. *16*, 3484–3488.

Collaborative Computational Project, Number 4. (1994). The CCP4 suite: programs for protein crystallography. Acta Crystallogr. D Biol. Crystallogr. *50*, 760–763.

Cooper, D.R., Boczek, T., Grelewska, K., Pinkowska, M., Sikorska, M., Zawadzki, M., and Derewenda, Z. (2007). Protein crystallization by surface entropy reduction: optimization of the SER strategy. Acta Crystallogr. D Biol. Crystallogr. *63*, 636–645.

Courchesne-Smith, C., Jang, S.H., Shi, Q., DeWille, J., Sasaki, G., and Kolattukudy, P.E. (1992). Cytoplasmic accumulation of a normally mitochondrial malonyl-CoA decarboxylase by the use of an alternate transcription start site. Arch. Biochem. Biophys. *298*, 576–586.

Cowtan, K. (2006). The Buccaneer software for automated model building. 1. Tracing protein chains. Acta Crystallogr. D Biol. Crystallogr. *62*, 1002–1011.

Cronan, J.E., Jr., and Waldrop, G.L. (2002). Multi-subunit acetyl-CoA carboxylases. Prog. Lipid Res. *41*, 407–435.

Doublié, S., Kapp, U., Aberg, A., Brown, K., Strub, K., and Cusack, S. (1996). Crystallization and preliminary X-ray analysis of the 9 kDa protein of the mouse signal recognition particle and the selenomethionyl-SRP9. FEBS Lett. *384*, 219–221.

Dyck, J.R.B., Hopkins, T.A., Bonnet, S., Michelakis, E.D., Young, M.E., Watanabe, M., Kawase, Y., Jishage, K.I., and Lopaschuk, G.D. (2006). Absence of malonyl coenzyme A decarboxylase in mice increases cardiac glucose oxidation and protects the heart from ischemic injury. Circulation *114*, 1721–1728.

Dyda, F., Klein, D.C., and Hickman, A.B. (2000). GCN5-related N-acetyltransferases: a structural overview. Annu. Rev. Biophys. Biomol. Struct. *29*, 81–103.

Emsley, P., and Cowtan, K.D. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr. D Biol. Crystallogr. *60*, 2126–2132.

Fantino, M. (2011). Role of lipids in the control of food intake. Curr. Opin. Clin. Nutr. Metab. Care *14*, 138–144.

Folmes, C.D.L., and Lopaschuk, G.D. (2007). Role of malonyl-CoA in heart disease and the hypothalamic control of obesity. Cardiovasc. Res. *73*, 278–287.

Fu, Z., Wang, M., Paschke, R., Rao, K.S., Frerman, F.E., and Kim, J.J. (2004). Crystal structure of human glutaryl-CoA dehydrogenase with and without an alternate substrate: structural bases of dehydrogenation and decarboxylation reactions. Biochemistry *43*, 9674–9684.

Gao, J., Waber, L., Bennett, M.J., Gibson, K.M., and Cohen, J.C. (1999). Cloning and mutational analysis of human malonyl-coenzyme A decarboxylase. J. Lipid Res. *40*, 178–182.

Goddard, T.D., Huang, C.C., and Ferrin, T.E. (2007). Visualizing density maps with UCSF Chimera. J. Struct. Biol. *157*, 281–287.

Goldschmidt, L., Cooper, D.R., Derewenda, Z.S., and Eisenberg, D. (2007). Toward rational protein crystallization: A Web server for the design of crystallizable protein variants. Protein Sci. *16*, 1569–1576.

Gu, L., Geders, T.W., Wang, B., Gerwick, W.H., Håkansson, K., Smith, J.L., and Sherman, D.H. (2007). GNAT-like strategy for polyketide chain initiation. Science *318*, 970–974.

Guillou, H., Zadravec, D., Martin, P.G.P., and Jacobsson, A. (2010). The key roles of elongases and desaturases in mammalian fatty acid metabolism: Insights from transgenic mice. Prog. Lipid Res. *49*, 186–199.

Holm, L., Kääriäinen, S., Rosenström, P., and Schenkel, A. (2008). Searching protein structure databases with DaliLite v.3. Bioinformatics *24*, 2780–2781.

Jansson, M., Li, Y.-C., Jendeberg, L., Anderson, S., Montelione, G.T., and Nilsson, B. (1996). High-level production of uniformly $^{15}$N- and $^{13}$C-enriched fusion proteins in Escherichia coli. J. Biomol. NMR *7*, 131–141.

Joly, E., Bendayan, M., Roduit, R., Saha, A.K., Ruderman, N.B., and Prentki, M. (2005). Malonyl-CoA decarboxylase is present in the cytosolic, mitochondrial and peroxisomal compartments of rat hepatocytes. FEBS Lett. *579*, 6581–6586.

Jung, J.S., Baek, D.J., Lee, G.Y., Kim, Y.S., and Oh, B.H. (2003). Crystallization and preliminary X-ray crystallographic analysis of malonyl-CoA decarboxylase from Rhizobium leguminosarum bv. trifolii. Acta Crystallogr. D Biol. Crystallogr. *59*, 166–167.

Kabsch, W. (2010). Integration, scaling, space-group assignment and post-refinement. Acta Crystallogr. D Biol. Crystallogr. *66*, 133–144.

Kim, Y.S., and Kolattukudy, P.E. (1978). Purification and properties of malonyl-CoA decarboxylase from rat liver mitochondria and its immunological comparison with the enzymes from rat brain, heart, and mammary gland. Arch. Biochem. Biophys. *190*, 234–246.

Kolattukudy, P.E., Poulose, A.J., and Kim, Y.S. (1981). Malonyl-CoA decarboxylase from avian, mammalian, and microbial sources. Methods Enzymol. *71*(Pt C), 150–163.

Kuhajda, F.P. (2006). Fatty acid synthase and cancer: new application of an old pathway. Cancer Res. *66*, 5977–5980.

Lane, M.D., Wolfgang, M., Cha, S.H., and Dai, Y. (2008). Regulation of food intake and energy expenditure by hypothalamic malonyl-CoA. Int. J. Obes. (Lond.) *32*(*Suppl 4*), S49–S54.

Laue, T.M., Shah, B.D., Ridgeway, T.M., and Pelletier, S.L. (1992). Computer-aided interpretation of analytical sedimentation data for proteins. In Analytical ultracentrifugation in biochemistry and polymer science, S.E. Harding, A.J. Rowe, and J.C. Horton, eds. (Cambridge: The Royal Society of Chemistry), pp. 90–125.

Lee, G.Y., Bahk, Y.Y., and Kim, Y.S. (2002). Rat malonyl-CoA decarboxylase; cloning, expression in E. coli and its biochemical characterization. J. Biochem. Mol. Biol. *35*, 213–219.

Liu, J., Montelione, G.T., and Rost, B. (2007). Novel leverage of structural genomics. Nat. Biotechnol. *25*, 849–851.

Lo, M.C., Wang, M., Kim, K.W., Busby, J., Yamane, H., Zondlo, J., Yuan, C., Young, S.W., and Xiao, S.H. (2008). A highly sensitive high-throughput luminescence assay for malonyl-CoA decarboxylase. Anal. Biochem. *376*, 122–130.

Lopaschuk, G.D., Ussher, J.R., and Jaswal, J.S. (2010). Targeting intermediary metabolism in the hypothalamus as a mechanism to regulate appetite. Pharmacol. Rev. *62*, 237–264.

Ludtke, S.J., Baldwin, P.R., and Chiu, W. (1999). EMAN: semiautomated software for high-resolution single-particle reconstructions. J. Struct. Biol. *128*, 82–97.

Malvagia, S., Papi, L., Morrone, A., Donati, M.A., Ciani, F., Pasquini, E., la Marca, G., Scholte, H.R., Genuardi, M., and Zammarchi, E. (2007). Fatal malonyl CoA decarboxylase deficiency due to maternal uniparental isodisomy of the telomeric end of chromosome 16. Ann. Hum. Genet. *71*, 705–712.

McGarry, J.D., and Brown, N.F. (1997). The mitochondrial carnitine palmitoyltransferase system. From concept to molecular analysis. Eur. J. Biochem. *244*, 1–14.

McRee, D.E. (1999). XtalView/Xfit—A versatile program for manipulating atomic coordinates and electron density. J. Struct. Biol. *125*, 156–165.

Nam, H.W., Lee, G.Y., and Kim, Y.S. (2006). Mass spectrometric identification of K210 essential for rat malonyl-CoA decarboxylase catalysis. J. Proteome Res. *5*, 1398–1406.

Neuwald, A.F., and Landsman, D. (1997). GCN5-related histone N-acetyltransferases belong to a diverse superfamily that includes the yeast SPT10 protein. Trends Biochem. Sci. *22*, 154–155.

Otwinowski, Z., and Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol. *276*, 307–326.

Ramsay, R.R., Gandour, R.D., and van der Leij, F.R. (2001). Molecular enzymology of carnitine transfer and transport. Biochim. Biophys. Acta *1546*, 21–43.

Sacksteder, K.A., Morrell, J.C., Wanders, R.J.A., Matalon, R., and Gould, S.J. (1999). MCD encodes peroxisomal and cytoplasmic forms of malonyl-CoA decarboxylase and is mutated in malonyl-CoA decarboxylase deficiency. J. Biol. Chem. *274*, 24461–24468.

Saggerson, D. (2008). Malonyl-CoA, a key signaling molecule in mammalian cells. Annu. Rev. Nutr. *28*, 253–272.

Salomons, G.S., Jakobs, C., Pope, L.L., Errami, A., Potter, M., Nowaczyk, M., Olpin, S., Manning, N., Raiman, J.A.J., Slade, T., et al. (2007). Clinical, enzymatic and molecular characterization of nine new patients with malonyl-coenzyme A decarboxylase deficiency. J. Inherit. Metab. Dis. *30*, 23–28.

Schuck, P. (2000). Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. Biophys. J. *78*, 1606–1619.

Sheldrick, G.M. (2008). A short history of SHELX. Acta Crystallogr. A *64*, 112–122.

Tang, H., Yan, Y., Feng, Z., de Jesus, R.K., Yang, L., Levorse, D.A., Owens, K.A., Akiyama, T.E., Bergeron, R., Castriota, G.A., et al. (2010). Design and synthesis of a new class of malonyl-CoA decarboxylase inhibitors with anti-obesity and anti-diabetic activities. Bioorg. Med. Chem. Lett. *20*, 6088–6092.

Terwilliger, T.C. (2003). SOLVE and RESOLVE: automated structure solution and density modification. Methods Enzymol. *374*, 22–37.

Tong, L. (2013). Structure and function of biotin-dependent carboxylases. Cell. Mol. Life Sci. *70*, 863–891.

Ussher, J.R., and Lopaschuk, G.D. (2009). Targeting malonyl CoA inhibition of mitochondrial fatty acid uptake as an approach to treat cardiac ischemia/reperfusion. Basic Res. Cardiol. *104*, 203–210.

Vagin, A.A., and Teplyakov, A. (2000). An approach to multi-copy search in molecular replacement. Acta Crystallogr. D Biol. Crystallogr. *56*, 1622–1624.

van Heel, M., Harauz, G., Orlova, E.V., Schmidt, R., and Schatz, M. (1996). A new generation of the IMAGIC image processing system. J. Struct. Biol. *116*, 17–24.

Vetting, M.W., S de Carvalho, L.P., Yu, M., Hegde, S.S., Magnet, S., Roderick, S.L., and Blanchard, J.S. (2005). Structure and functions of the GNAT superfamily of acetyltransferases. Arch. Biochem. Biophys. *433*, 212–226.

Vetting, M.W., Errey, J.C., and Blanchard, J.S. (2008). Rv0802c from Mycobacterium tuberculosis: the first structure of a succinyltransferase with the GNAT fold. Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. *64*, 978–985.

Vonrhein, C., Blanc, E., Roversi, P., and Bricogne, G. (2007). Automated structure solution with autoSHARP. Methods Mol. Biol. *364*, 215–230.

Wakil, S.J., Stoops, J.K., and Joshi, V.C. (1983). Fatty acid synthesis and its regulation. Annu. Rev. Biochem. *52*, 537–579.

Wallace, D.M., Haramura, M., Cheng, J.-F., Arrhenius, T., and Nadzan, A.M. (2007). Novel trifluoroacetophenone derivatives as malonyl-CoA decarboxylase inhibitors. Bioorg. Med. Chem. Lett. *17*, 1127–1130.

Wightman, P.J., Santer, R., Ribes, A., Dougherty, F., McGill, N., Thorburn, D.R., and FitzPatrick, D.R. (2003). MLYCD mutation analysis: evidence for protein mistargeting as a cause of MLYCD deficiency. Hum. Mutat. *22*, 288–300.

Wolfgang, M.J., and Lane, M.D. (2008). Hypothalamic malonyl-coenzyme A and the control of energy balance. Mol. Endocrinol. *22*, 2012–2020.

Xue, J., Peng, J., Zhou, M., Zhong, L., Yin, F., Liang, D., and Wu, L. (2012). Novel compound heterozygous mutation of MLYCD in a Chinese patient with malonic aciduria. Mol. Genet. Metab. *105*, 79–83.

Zammit, V.A. (1999). The malonyl-CoA-long-chain acyl-CoA axis in the maintenance of mammalian cell function. Biochem. J. *343*, 505–515.

Zhou, D., Yuen, P., Chu, D., Thon, V., McConnell, S., Brown, S., Tsang, A., Pena, M., Russell, A., Cheng, J.-F., et al. (2004). Expression, purification, and characterization of human malonyl-CoA decarboxylase. Protein Expr. Purif. *34*, 261–269.

Zhou, W., Tu, Y., Simpson, P.J., and Kuhajda, F.P. (2009). Malonyl-CoA decarboxylase inhibition is selectively cytotoxic to human breast cancer cells. Oncogene *28*, 2979–2987.

Biochemical Society TRANSACTIONS

# Electron microscopy studies of Type III CRISPR machines in *Sulfolobus solfataricus*

**Giuseppe Cannone*, Mariam Webber-Birungi* and Laura Spagnolo*[1]**

*Institute of Structural Molecular Biology and Centre for Science at Extreme Conditions, University of Edinburgh, Edinburgh EH9 3JR, U.K.

## Abstract

**The CRISPR (clustered regularly interspaced short palindromic repeats) system is an adaptive immune system that targets viruses and other mobile genetic elements in bacteria and archaea. Cells store information of past infections in their genome in repeat–spacer arrays. After transcription, these arrays are processed into unit-length crRNA (CRISPR RNA) that is loaded into effector complexes encoded by *Cas* (CRISPR-associated) genes. CRISPR–Cas complexes target invading nucleic acid for degradation. CRISPR effector complexes have been classified into three main types (I–III). Type III effector complexes share the Cas10 subunit. In the present paper, we discuss the structures of the two Type III effector complexes from *Sulfolobus solfataricus*, SsoCSM (subtype III-A) and SsoCMR (subtype III-B), obtained by electron microscopy and single particle analysis. We also compare these structures with Cascade (CRISPR-associated complex for antiviral defence) and with the RecA nucleoprotein.**

## Introduction

In prokaryotes, CRISPRs (clustered regularly interspaced short palindromic repeats) are involved in an interference pathway that protects cells from bacteriophages and viruses. CRISPR sequences confer an adaptive heritable trace of past infections and express crRNAs (CRISPR RNAs), short RNAs that target 'non-self' nucleic acids. Cas (CRISPR-associated) proteins are integral players of the 'prokaryotic immune system' termed CRISPR–Cas defence. The Cas9 endonuclease CRISPR–Cas system has recently emerged as a powerful tool for genome editing in various cells and organisms.

CRISPR–Cas complexes have been extensively studied in recent times with structural biology methods to gain an insight into their molecular mechanism [1]. In Type III systems, Csm and Cmr proteins are known to form functional complexes involved in DNA and RNA targeting respectively. In the present paper, we discuss the structures of two archaeal interference complexes from *Sulfolobus solfataricus*, SsoCSM and SsoCMR, as determined by electron microscopy and single particle analysis. We highlight analogies and differences with the RecA structure, as well as with other CRISPR–Cas proteins for which structural information is available.

## CRISPR–Cas complexes

The CRISPR–Cas prokaryotic defence consists of a multistep process whereby foreign nucleic acids are first recognized as being non-self and incorporated into the host genome between short DNA repeats. These small fragments, in conjunction with host Cas proteins, are then used to recognize and destroy foreign nucleic acids. The dual tracrRNA (transactivating crRNA)–crRNA programmable Cas9 endonuclease of the Type II CRISPR–Cas system has proved to be an effective genome-editing tool in different cells and organisms [2–12].

CRISPR–Cas systems are classified into Type I, Type II and Type III based on their phylogeny, sequence, locus organization and content of the CRISPRs and associated *Cas* genes [13–15]. The protein Cas3 is a signature of Type I systems, Cas9 is a signature of Type II systems, and Cas10 is a signature of Type III systems. These types are further divided into ten subtypes. Two Type III CRISPR–Cas complexes have been identified in the archaeon *S. solfataricus*. SsoCSM is a subtype III-A complex directed towards DNA. SsoCMR is a subtype III-B complex directed towards RNA.

## Electron microscopy studies of the SsoCSM complex

The Sso subtype III-A effector complex, also known as the CSM complex, is associated with crRNA generated by cleavage of pre-crRNA following 5′- and 3′-end processing of pre-crRNA by Cas6 and an unknown nuclease [16]. The requirement for a mismatch region at the boundary of the repeat–spacer sequence ensures that the CRISPR locus in the host genome is not cleaved by subtype III-A systems [17].

We solved the structure of the SsoCSM–RNA complex with electron microscopy coupled to single particle analysis [18] (Figure 1A). The complex exhibits an elongated structure formed by two intertwined filaments connected at one end by a large base. Direct comparison with the bacterial Cascade (CRISPR-associated complex for antiviral defence) complex [19] reveals crucial analogies to the anatomy of Type I

**Key words:** clustered regularly interspaced short palindromic repeats (CRISPR)–CRISPR-associated (Cas), CRISPR RNA (crRNA), RecA, RNA polymerase, Type III CRISPR.

**Abbreviations used:** Cas, CRISPR-associated; Cascade, CRISPR-associated complex for antiviral defence; CRISPR, clustered regularly short interspaced palindromic repeats; crRNA, CRISPR RNA; RAMP, repeat-associated mysterious protein; Sso, *Sulfolobus solfataricus*.

[1]To whom correspondence should be addressed (email laura.spagnolo@ed.ac.uk).

**Figure 1** | **Three-dimensional structure of the SsoCSM complex**

(**A**) Four orthogonal views of the complex. (**B**) Cascade backbone (blue) fitted into the CSM complex (grey). Figures were prepared using the Chimera software [30].

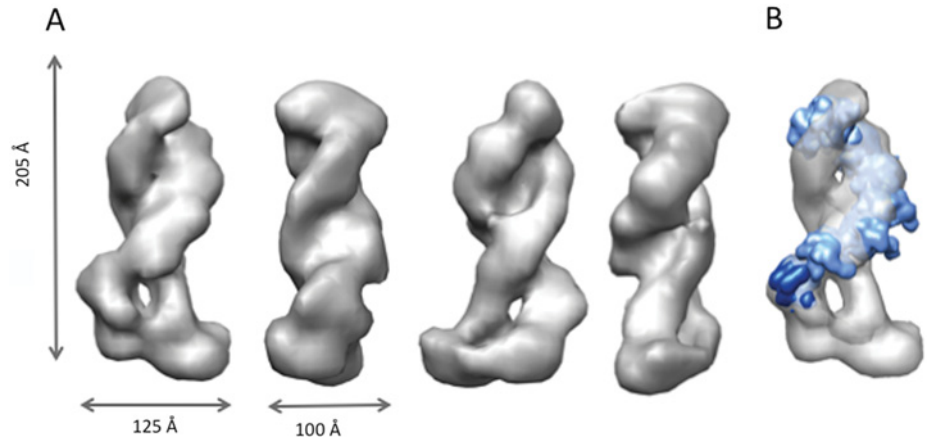
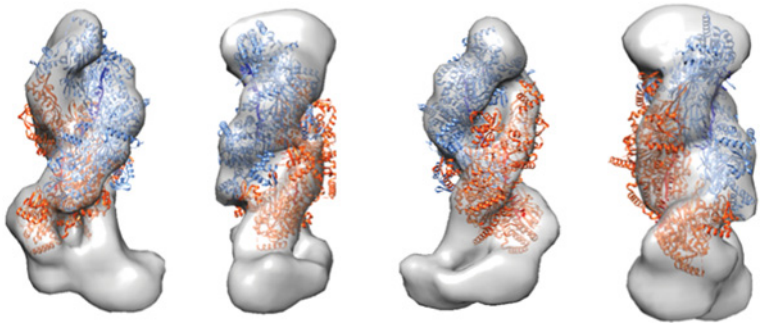



**Figure 2** | **Fitting RecA into the map of the SsoCSM complex**

RecA (PDB code 3CMW) was fitted into the electron microscopy map of the SsoCMR complex using the Chimera software [30].



systems (Figure 1B). In particular, the backbone formed by six Cas7-proximal domains and the Cas5 subunit from the bacterial Cascade complex can be accurately superimposed to the RAMP (repeat-associated mysterious protein) subunits (Sso1427, four copies of Sso1426 and Sso1432) in CSM. Both pitch and height of the Cascade backbone are identical with that of CSM. SsoCSM is slightly longer than the bacterial Cascade (205 Å compared with 190 Å; where 1 Å = 0.1 nm). The larger diameter of the major backbone is consistent with the presence of bound crRNA, probably in an orientation analogous to the one observed in Cascade. Consistent with the distinct structures of Cse1 and Cmr2, which are the large subunits of the Type I and Type III complexes, the bases of the two complexes are not structurally similar.

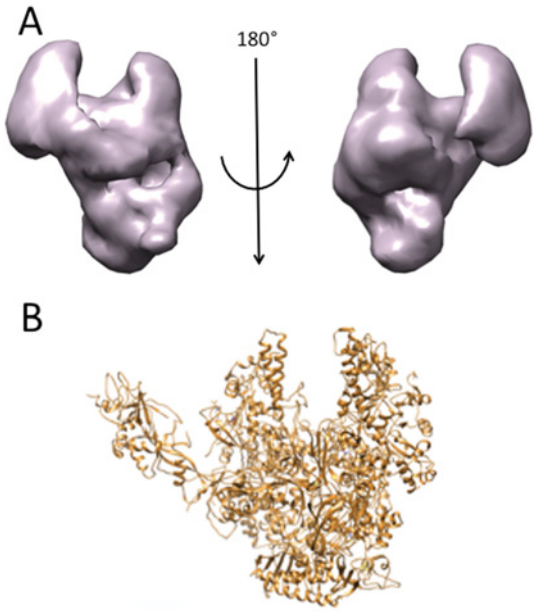
Structural similarities between the Cascade–RNA complex and the RecA–ssDNA nucleoprotein [20] have been highlighted previously [21,22]. This comparison is very relevant with respect to the RecA mechanism of recognition of homologous dsDNA and strand exchange, where the protein wraps around the nucleic acid, in keeping an overlap of the two helical axes. We have therefore tested how the RecA crystallographic structure (PDB code 3CMW [20]) fits into the CSM map (Figure 2). In this case, we fitted both protein chains (A and C) from the PDB file, each of which is superimposable with one of the two coiled filaments in CSM. Pitch, height and filament width are consistent between RecA and the upper part of CSM. This could suggest a possible RNA/DNA-scanning and -recognition interface in correspondence of the crevice between CSM filaments. In fact, this cleft could well accommodate in length and width a 38 bp dsDNA target sequence, such as the Locus A spacer 26-abundant crRNA sequence identified in the CSM complex by deep sequencing [18].

## Electron microscopy studies of the SsoCMR complex

The electron microscopy structure of the CMR complex [23] (Figure 3) has no obvious similarity to the 'seahorse' structure of *Escherichia coli* Cascade [24] or to the RecA–ssDNA nucleoprotein [20]. CMR is structurally rather reminiscent of

**Figure 3 | Three-dimensional reconstruction of SsoCMR and comparison with RNA polymerase**

(**A**) Surface representation of two faces of CMR. (**B**) Ribbon representation of the Sso RNA polymerase (PDB code 2PMZ). Figures were prepared using Chimera software [30].




RNA polymerase [25] (Figure 3B), as initially hypothesized with sequence analysis tools. In particular, its upper half is organized in a 'claw' region, which could be a dsRNA-binding cleft. This is in line with Cmr2 harbouring the active site of the CMR complex. The Cmr2–Cmr3–Cmr7 subcomplex, which contains no bound crRNA, has a deeper cleft in comparison with the full complex. The lack of crRNA in the Cmr2–Cmr3–Cmr7 subcomplex fits with a presumed role for the RAMP-containing Cmr subunits (Cmr1, Cmr4 and Cmr6) in RNA binding [14]. Compared with Cmr2–Cmr3–Cmr7, the additional Cmr subunits are distributed mainly at the front and at the tail of the complete CMR complex. Identification of the path of RNA in the CMR structure remains elusive; however, it will be vital to elucidate the molecular mechanism and organization of this complex.

Recent crystallographic work on CMR subunits has unveiled the structure of the Cmr7 dimer (Sso1986) [23] and of a truncation mutant of Cmr2 in isolation [26,27], as well as in complex with Cmr3 [28]. We tested fitting the related PDB files into the envelope of electron density to provide a model of how each of these subunits is arranged relative to each other in the Cmr2–Cmr3–Cmr7 subcomplex, as well as in the full CMR complex. Volumetric analysis of the sum of the volumes is consistent with the Cmr2–Cmr3–Cmr7 complex stoichiometry determined by densitometry (i.e. 1:1:6) [23]. Unfortunately, it was not possible to unequivocally fit the PDB files in the electron microscopy map because we lacked strong structural features that could guide this analysis. Positional mapping of CMR, which would be instrumental to understanding the interference mechanism for this complex, therefore remains elusive.

## Comparison between SsoCSM and SsoCMR

The electron microscopy structure of the subtype III-B (CMR) structure appears very different from that of the subtype III-A complex, despite the fact that they share much clearer homology than either does with Cascade. The 'body' of the CMR complex is composed of a number of RAMP-domain proteins (Cmr1, Cmr4, Cmr5 and Cmr6) that are assumed to bind RNA. However, they are not obviously arranged in the helical conformation seen for the Type I and subtype III-A complexes, but appear to form a more compact structure [29]. This may reflect the fact that CMR targets flexible RNA substrates, rather than rigid helical dsDNA. The mechanism of molecular recognition for these effectors could therefore be expected to differ fundamentally and be reflected in their distinct structures.

## Conclusions

Structural studies of CRISPR–Cas assemblies has proven a very productive field in recent times [1]. Structural electron microscopy is a powerful technique to identify and analyse similarities and differences between the large complexes classified in each type, therefore providing hints for future studies aimed at deciphering the molecular mechanisms involved in every process. A key feature that many CRISPR–Cas complexes have in common is the presence of a backbone (sometimes compared with a 'spine' in the anatomy of the complex), which binds the RNA component of the ribonucleoprotein. Higher-resolution comparative structural studies will help to unravel how these backbones differ and how these differences are related to each functional mechanism. CRISPR–Cas systems have been utilized for efficient genome editing [2–12]. Understanding their structure at the molecular level bears a strong potential to optimize their use for applications such as gene therapy.

## Acknowledgements

We thank Malcolm F. White and members of his group for constructive discussions.

## Funding

This work was supported by the Royal Society [grant numbers RG2009/R1; R41318] and the Biotechnology and Biological Sciences Research Council [grant number BB/J005673/1] (to L.S.), and the Darwin Trust of Edinburgh (to G.C.). The Electron Microscopy Facility at Edinburgh is supported by the Scottish Alliance for Life Sciences and the Wellcome Trust [grant number WT087658MA].

## References


1 Reeks, J., Naismith, J.H. and White, M.F. (2013) CRISPR interference: a structural perspective. Biochem. J. **453**, 155–166
2 Gratz, S.J., Cummings, A.M., Nguyen, J.N., Hamm, D.C., Donohue, L.K., Harrison, M.M., Wildonger, J. and O'Connor-Giles, K.M. (2013) Genome engineering of *Drosophila* with the CRISPR RNA-guided Cas9 nuclease. Genetics **194**, 1029–1035
3 Gaj, T., Gersbach, C.A. and Barbas, III, C.F. (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol. **31**, 397–405
4 Wang, H., Yang, H., Shivalila, C.S., Dawlaty, M.M., Cheng, A.W., Zhang, F. and Jaenisch, R. (2013) One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. Cell **153**, 910–918
5 DiCarlo, J.E., Norville, J.E., Mali, P., Rios, X., Aach, J. and Church, G.M. (2013) Genome engineering in *Saccharomyces cerevisiae* using CRISPR–Cas systems. Nucleic Acids Res. **41**, 4336–4343
6 Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A. et al. (2013) Multiplex genome engineering using CRISPR/Cas systems. Science **339**, 819–823
7 Jiang, W., Bikard, D., Cox, D., Zhang, F. and Marraffini, L.A. (2013) RNA-guided editing of bacterial genomes using CRISPR–Cas systems. Nat. Biotechnol. **31**, 233–239
8 Hwang, W.Y., Fu, Y., Reyon, D., Maeder, M.L., Tsai, S.Q., Sander, J.D., Peterson, R.T., Yeh, J.R. and Joung, J.K. (2013) Efficient genome editing in zebrafish using a CRISPR–Cas system. Nat. Biotechnol. **31**, 227–229
9 Friedland, A.E., Tzur, Y.B., Esvelt, K.M., Colaiacovo, M.P., Church, G.M. and Calarco, J.A. (2013) Heritable genome editing in *C. elegans* via a CRISPR–Cas9 system. Nat. Methods **10**, 741–743
10 Burgess, D.J. (2013) Technology: a CRISPR genome-editing tool. Nat. Rev. Genet. **14**, 80
11 Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A. and Charpentier, E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science **337**, 816–821
12 Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E. and Church, G.M. (2013) RNA-guided human genome engineering via Cas9. Science **339**, 823–826
13 Bhaya, D., Davison, M. and Barrangou, R. (2011) CRISPR–Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. Annu. Rev. Genet. **45**, 273–297
14 Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J., Wolf, Y.I., Yakunin, A.F. et al. (2011) Evolution and classification of the CRISPR-Cas systems. Nat. Rev. Microbiol. **9**, 467–477
15 Makarova, K.S., Aravind, L., Wolf, Y.I. and Koonin, E.V. (2011) Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR–Cas systems. Biol. Direct **6**, 38
16 Hatoum-Aslan, A., Maniv, I. and Marraffini, L.A. (2011) Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site. Proc. Natl. Acad. Sci. U.S.A. **108**, 21218–21222
17 Marraffini, L.A. and Sontheimer, E.J. (2010) Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature **463**, 568–571
18 Rouillon, C., Zhou, M., Zhang, J., Politis, A., Beilsten, V., Cannone, G., Graham, S., Robinson, C., Spagnolo, L. and White, M. (2013) Structure of the CRISPR surveillance complex CSM reveals key similarities with Cascade. Mol. Cell **52**, 124–134
19 Wiedenheft, B., Lander, G.C., Zhou, K., Jore, M.M., Brouns, S.J., van der Oost, J., Doudna, J.A. and Nogales, E. (2011) Structures of the RNA-guided surveillance complex from a bacterial immune system. Nature **477**, 486–489
20 Chen, Z., Yang, H. and Pavletich, N.P. (2008) Mechanism of homologous recombination from the RecA-ssDNA/dsDNA structures. Nature **453**, 489–484
21 Lintner, N.G., Kerou, M., Brumfield, S.K., Graham, S., Liu, H., Naismith, J.H., Sdano, M., Peng, N., She, Q., Copie, V. et al. (2011) Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). J. Biol. Chem. **286**, 21643–21656
22 Sorek, R., Lawrence, C.M. and Wiedenheft, B. (2013) CRISPR-mediated adaptive immune systems in bacteria and archaea. Annu. Rev. Biochem. **82**, 237–266
23 Zhang, J., Rouillon, C., Kerou, M., Reeks, J., Brugger, K., Graham, S., Reimann, J., Cannone, G., Liu, H., Albers, S.V. et al. (2012) Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. Mol. Cell **45**, 303–313
24 Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, U., Wurm, R., Wagner, R. et al. (2011) Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat. Struct. Mol. Biol. **18**, 529–536
25 Hirata, A., Klein, B.J. and Murakami, K.S. (2008) The X-ray crystal structure of RNA polymerase from archaea. Nature **451**, 851–854
26 Cocozaki, A.I., Ramia, N.F., Shao, Y., Hale, C.R., Terns, R.M., Terns, M.P. and Li, H. (2012) Structure of the Cmr2 subunit of the CRISPR–Cas RNA silencing complex. Structure **20**, 545–553
27 Zhu, X. and Ye, K. (2012) Crystal structure of Cmr2 suggests a nucleotide cyclase-related enzyme in Type III CRISPR–Cas systems. FEBS Lett. **586**, 939–945
28 Shao, Y., Cocozaki, A.I., Ramia, N.F., Terns, R.M., Terns, M.P. and Li, H. (2013) Structure of the Cmr2–Cmr3 subcomplex of the Cmr RNA silencing complex. Structure **21**, 376–384
29 Zhang, J., Rouillon, C., Kerou, M., Reeks, J., Brugger, K., Graham, S., Reimann, J., Cannone, G., Liu, H., Albers, S.V. et al. (2012) Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. Mol. Cell **45**, 303–313
30 Goddard, T.D., Huang, C.C. and Ferrin, T.E. (2007) Visualizing density maps with UCSF Chimera. J. Struct. Biol. **157**, 281–287


Received 4 September 2013
doi:10.1042/BST20130166