The H-subunit of the restriction endonuclease CglI contains a prototype DEAD-Z1 helicase-like motor

Paulius Toliusis1,†, Giedre Tamulaitiene1,‡, Rokas Grigaitis1, Donata Tumauskaite1, Arunas Silanskas1, Elena Manakova1, Česlovas Venclovas1, Mark D. Szczelkun2, Virginijus Siksnys1,† and Mindaugas Zaremba1,‡

1Institute of Biotechnology, Life Sciences Center, Vilnius University, Saulėtekio al. 7, LT-10257 Vilnius, Lithuania and 2DNA–Protein Interactions Unit, School of Biochemistry, Biomedical Sciences Building, University of Bristol, Bristol BS8 1TD, UK

Received December 07, 2017; Revised January 18, 2018; Editorial Decision February 03, 2018; Accepted February 08, 2018

INTRODUCTION

Helicases are a large class of nucleic acid enzymes with a wide range of biological functions (1). The classical activity for which they are named is the NTP hydrolysis-dependent strand separation of duplex polynucleotides. The single-strand products are important intermediates in processes such as DNA replication and repair, and RNA metabolism. Additionally, helicase enzymes have evolved for nucleic acid processing activities, including dsDNA translocation (e.g. in the chromatin remodelling enzymes (2)), or acting as molecular switches (e.g. in the Type III restriction enzymes (3)). In these enzymes the domains are referred to as ‘helicase-like’, to highlight the lack of unwinding activity (4). Here, we examined the structural and mechanistic features of the H-subunit of the Type I restriction enzyme CglI (5,6), which represents an alternative use of helicase-like activity.

Helicase and helicase-like enzymes are members of the P-loop superfamily of NTPases which are characterised by the Walker A and Walker B catalytic motifs. They can be further identified and classified based on different structural features which help define their function (7–12): (i) helicases are grouped by homologous amino acid sequences into six superfamilies, SF1 – SF6; (ii) the core structural fold of all helicases is a RecA-like domain or the related AAA+ domain, comprising an α–β structure which folds into an antiparallel β-sheet sandwiched by the α-helices. Helicase motifs are on the intervening loops. Additional unique domains can be identified which provide additional activities, such as dsDNA binding; (iii) helicases can be active as either monomers, or higher order assemblies, typically hexameric rings. In all cases, the catalytic NTP binds at the interface between two RecA/AAA+ domains. For the monomeric helicases, a pair of domains are found in one polypeptide and are named N-core (containing the Walker A and B) and C-core. For hexameric helicases, the interface is formed between an adjacent pair of protomers.

Restriction enzymes (REs) such as CglI evolved as prokaryotic immune systems to restrict bacteriophage

To whom correspondence should be addressed. Tel: +370 5 2234357; Fax: +370 5 2234367; Email: zare@ibt.lt
Correspondence may also be addressed to Virginijus Siksnys. Tel: +370-5-2234359; Fax: +370-5-2234367; Email: siksnys@ibt.lt
†These authors contributed equally to this work as first authors.
Present address: Rokas Grigaitis, Institut für Biochemie, ETH Zürich, HPM D 6.1.Otto-Stern-Weg 3, 8093 Zürich, Switzerland.

© The Author(s) 2018. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
dent upon ATP hydrolysis (14); these enzymes are classi-
fied as Type I, ISP or III (15,16). The Type I RE CglI com-
icates its target sequence 5′-GCCGC-3′ main and a C-terminal B3-like domain which binds the spe-
ific site motif H104 and K106 are coloured red). B3-like domains (coloured blue and light blue) dimerise to form single active site (nuclease active site). The 632 aa H-subunit has an N-terminal phospholipase D (PLD) do-
main (PLD-superfamily nuclease and B3-like DNA binding do-
main boundaries. (B) A possible model of the R2H2.CglI complex (6,79). The R.CglI dimer is presented by the structure of the homologous R.NgoAVII (PDB ID 4RCT) bound to the cognate DNA via the B3 domains (PDB ID 4RD5). Two R.NgoAVII PLD domains from separate subunits (coloured blue and light blue) dimerise to form single active site (nucleic acid site residues H104 and K106 are coloured red). B3-like domains (coloured green and light green) are found in plant transcription factors and recog-
nise the 5′-GCCGC-3′ target. DNA is shown as black ribbon. H.CglI con-
tains the DEAD, Z1 and C-terminal domains (separate subunits are white or light grey).

Figure 1. The restriction endonuclease CglI from Corynebacterium glutamicum. (A) Domain organisation of R- and H.CglI proteins. R.CglI (NCgl1704) has PLD-superfamily nuclease and B3-like DNA binding domains. H.CglI (NCgl1705) has a predicted SF2 helicase/ATPase do-
main (DEAD) linked to uncharacterized Z1-superfamily and C-terminal domains. Numbers represent predicted amino acid positions of domain boundaries. (B) A possible model of the R2H2.CglI complex (6,79). The R.CglI dimer is presented by the structure of the homologous R.NgoAVII (PDB ID 4RCT) bound to the cognate DNA via the B3 domains (PDB ID 4RD5). Two R.NgoAVII PLD domains from separate subunits (coloured blue and light blue) dimerise to form single active site (nucleic acid site residues H104 and K106 are coloured red). B3-like domains (coloured green and light green) are found in plant transcription factors and recog-
nise the 5′-GCCGC-3′ target. DNA is shown as black ribbon. H.CglI con-
tains the DEAD, Z1 and C-terminal domains (separate subunits are white or light grey).

growth or transformation by other genetic elements, by cleaving the target DNA following sequence specific recognition (13). At least a third of REs have SF2 helicase-like domains and have DNA cleavage activity that is dependent upon ATP hydrolysis (14); these enzymes are classi-
fied as Type I, ISP or III (15,16). The Type I RE CglI comprises two subunits, R and H (Figure 1A) (5,6). The 358 aa R-subunit has an N-terminal phospholipase D (PLD) do-
main and a C-terminal B3-like domain which binds the spe-
cific target sequence 5′-GCCGC-3′. The 632 aa H-subunit comprises an N-terminal 229 aa helicase domain (named ‘DEAD’ after the Walker B sequence), an uncharacterised 234 aa Z1-superfamily domain (17), and an uncharacterized 169 aa C-terminal domain with a lack of primary sequence similarity to proteins of known structure. CglI is active as a heterotetrameric R2H2 complex, formed of two dimers R2 and H2, where dimerization occurs via the PLD domains (to generate a dimeric nuclease with a single active site), and via the Z1 domains, respectively (Figure 1B). Following site-specific DNA recognition by the B3 domain(s), exten-
sive nucleotide hydrolysis (~170 ATP/s) is coupled to step-
wise DNA translocation, preferentially in a downstream di-
rection from the target, although upstream translocation is also possible (5). Unusually, translocating motors can transfer onto a new DNA substrate with low efficiency. Upon colli-
usion with another enzyme at a target site, DNA cleavage occurs. The most efficient cleavage activity is due to ‘rear-
end’, so circular DNA and linear DNA with sites in direct repeat (i.e. head-to-tail) are the best cleavage substrates.

The elevated ATP hydrolysis rate and roadblock displacement activities of CglI are consistent with a dsDNA translo-
cating motor like other Type I and ISP REs (4). The DEAD do-
main was identified based on the homology of conserved SF2 motifs, including the characteristic Walker A and B (18,19). However, at 229 aa (Figure 1A), the domain is appreciably smaller than the SF2 helicase-like motors of other Type I and ISP enzymes, which are typically >400 aa monomers comprising both N- and C-core RecA-like sub-
domains (4,7). This is also the arrangement observed in all other SF2 enzymes (20). The DEAD domain of CglI has a size more consistent with a single RecA fold. Key to under-
standing the mechanism of CglI is therefore an understanding of the structural arrangement of the unusual DEAD do-
main.

Based on our simple model for the R2H2 complex (Fig-
ure 1B) (6), three structural arrangements for the helicase-
like motor can be suggested: (i) The DEAD domain alone has ATP binding and hydrolysis activity, meaning there are two independent ATPase sites per R2H2 complex. This model is disfavoured because all helicases studied to-date, regardless of superfamily, use two RecA folds to generate an NTP binding interface. Moreover, it is the opening-closing motions of the domains during the ATPase cycle that are fundamental to mecanochemochemical coupling. (ii) The two DEAD domains in the complex dimerise to produce one ATP binding interface. This would allow opening-closing motions of the domains that could be coupled to dsDNA translocation. This model is disfavoured because dimerization between isolated DEAD domains has not been ob-
erved (6). Additionally, this arrangement would be unusual for a SF2 enzyme, and interactions between identical do-
 mains in SF3-SF6 is via multimerization, not dimerization. (iii) The DEAD domain is only one half of the helicase-
like motor, and another domain provides the second RecA-
fold. Given the presence of the Walker A and B motifs, the DEAD domain can be defined as the N-core. The most likely candidate for the C-core would be the adjacent 234 aa Z1 domain. This arrangement would be more like other SF2 enzymes where the N- and C-core are contiguous. Therefore, we propose the R2H2 complex would contain two sepa-
rate DEAD-Z1 helicase-like motors.

To examine whether CglI employs a prototype DEAD-
Z1 helicase-like motor and to further expand on our over-
all understanding of the mechanism of this class of REs, we employed a combination of structural approaches (includ-
ing X-ray crystallography, small-angle X-ray scattering and molecular modelling) and biochemical analysis (DNA- and ATP-binding, ATP hydrolysis and DNA cleavage) with protein mutagenesis. Our results indicate that DEAD-Z1 do in-
deed together form a unique ATP- and DNA-binding heli-
case fold with specific helicase motifs (Loop 163–172, H/H, Motif V and Motif VI), while a crystal structure of the C-
terminal domain revealed a novel fold with DNA-binding activity. Despite the independent activities of the domains,
we could not reconstitute ATPase or nuclease activity from any combination of the purified domains. We additionally found that when even one of the C-terminal domains in the complex was deleted, DNA cleavage activity was reduced suggesting that both H-subunits (i.e. two motors) cooperate in the process. Finally, using SAXS we are able to determine in more detail the relative arrangement of the H-subunit dimer.

**MATERIALS AND METHODS**

**Expression and purification of CglII proteins**

The CglII proteins, their mutants and the isolated domains and domain fusions were expressed and purified as described earlier (6). Se-Met modified H.CglII C-terminal domain was prepared by inhibiting the methionine synthesis pathway and expressing the protein in the presence of D,L-selenomethionine (Sigma) following the published procedure (21). The identity of the purified proteins was confirmed by mass spectrometry. All protein concentrations used in this work are shown in terms of monomer.

**Mutagenesis**

Mutants of the H.CglII protein and its domains were produced by QuikChange Site-Directed Mutagenesis (22). Sequencing of the entire gene for each mutant confirmed that only the desired mutations had been introduced.

**Crystallization and structure determination**

The unlabelled and Se-Met labelled C-terminal domains were concentrated to 15 mg/ml in buffer (10 mM Tris–HCl (pH 7.5 at 25°C), 300 mM KCl, 1 mM EDTA, 1 mM DTT). Crystals were grown using the sitting drop vapor diffusion method at 19°C by mixing 0.5 μl of the protein solution with 0.5 μl of the crystallization buffer (0.1 M Tris–HCl (pH 8.5 at 25°C), 7–14% (v/v) Tacsimate, 25–30% (w/v) PEG3350). Crystals belonging to the space group P6122 appeared after few days. The X-ray diffraction datasets for Se-Met modified protein crystals were collected at the EMBL-Hamburg automated crystallography Table S1.

**Crystals were grown using the sitting drop vapor diffusion procedure (6). Se-Met modified H.CglII C-terminal domain fusion was expressed and purified as described earlier (6). The CglII proteins, their mutants and the isolated domains and domain fusions were expressed and purified as described earlier (6). The crystal contacts were determined in more detail the relative arrangement of the H-subunit dimer.**

**Expression and purification of CglII proteins**

The CglII proteins, their mutants and the isolated domains and domain fusions were expressed and purified as described earlier (6). Se-Met modified H.CglII C-terminal domain was prepared by inhibiting the methionine synthesis pathway and expressing the protein in the presence of D,L-selenomethionine (Sigma) following the published procedure (21). The identity of the purified proteins was confirmed by mass spectrometry. All protein concentrations used in this work are shown in terms of monomer.

**MATERIALS AND METHODS**

**Expression and purification of CglII proteins**

The CglII proteins, their mutants and the isolated domains and domain fusions were expressed and purified as described earlier (6). Se-Met modified H.CglII C-terminal domain was prepared by inhibiting the methionine synthesis pathway and expressing the protein in the presence of D,L-selenomethionine (Sigma) following the published procedure (21). The identity of the purified proteins was confirmed by mass spectrometry. All protein concentrations used in this work are shown in terms of monomer.

**Mutagenesis**

Mutants of the H.CglII protein and its domains were produced by QuikChange Site-Directed Mutagenesis (22). Sequencing of the entire gene for each mutant confirmed that only the desired mutations had been introduced.

**Crystallization and structure determination**

The unlabelled and Se-Met labelled C-terminal domains were concentrated to 15 mg/ml in buffer (10 mM Tris–HCl (pH 7.5 at 25°C), 300 mM KCl, 1 mM EDTA, 1 mM DTT). Crystals were grown using the sitting drop vapor diffusion method at 19°C by mixing 0.5 μl of the protein solution with 0.5 μl of the crystallization buffer (0.1 M Tris–HCl (pH 8.5 at 25°C), 7–14% (v/v) Tacsimate, 25–30% (w/v) PEG3350). Crystals belonging to the space group P6122 appeared after few days. The X-ray diffraction datasets for Se-Met modified protein crystals were collected at the EMBL-Hamburg automated crys-
out at room temperature using $\lambda_{ex} = 356$ nm and $\lambda_{em} = 440$ nm. Dissociation constant values ($K_d$) for Mant-ATP binding were calculated from 3–5 consecutive measurements by fitting binding data to the hyperbolic Equation (2).

### ATPase assay

ATPase reactions were performed at 37°C in reaction buffer (33 mM Tris-acetate (pH 7.0 at 25°C), 10 mM Mg-acetate, 0.1 mg/ml BSA, 1 mM DTT, 150 mM K-acetate) supplemented with 0.62 nM phage CglI DNA (contains 181 CglI site) and 2 mM ATP. Reactions were initiated by adding 10 nM H.CglI or its domain(s) and 200 nM R.CglI. Aliquots were removed at fixed time intervals and the reactions were stopped by adding EDTA to a final concentration of 9.5 mM. Phosphate accumulation rate ($v$, $\mu$M × min$^{-1}$) was measured using Malachite green phosphate assay kit (BioAssay Systems), which is used to measure the ATP hydrolysis rate by detection of liberated free phosphate, following the manufacturer's protocol. The relationship between the absorbance and phosphate concentration was described (5).

### DNA binding assay

DNA binding by the CglI proteins was analysed by the electrophoretic mobility shift assay (EMSA) using a 25 bp non-cognate oligoduplex (5′-AGATCGAAGACCGTAGCACTGGGCT-3′/5′-TCTAGCCTTCTGCATCGTAGCCCA-5′). The DNA was labelled at the 5′ end on one strand with $^{32}$P DNA (final concentration 100 nM) was incubated with H.CglI proteins or domains (final concentrations from 50 to 5000 nM) for 20 min in 10 μl of the binding buffer containing 30 mM MES-His (pH 6.0 at 25°C), 1 mM DTT, 0.1 mg/ml BSA and 10% (v/v) glycerol at room temperature. Free DNA and protein–DNA complexes were separated by electrophoresis using 8% (w/v) acrylamide gels (29:1 acrylamide/bisacrylamide in 30 mM MES-His, pH 6.0 at 25°C). Samples were separated at room temperature for 1 h at ~6 V/cm. Radiolabelled DNA was detected and quantified using the Fujifilm FLA-5100 fluorescent image analyzer (Fujifilm, Tokyo, Japan) and the OptiQuant 3.0 software (Packard Instrument) to determine the volume of each band, taking background readings into account. Apparent $K_d$ values for protein–DNA binding were calculated from three independent reactions by fitting the binding data to Equation (1).

### DNA cleavage assay

DNA cleavage activity was tested using 10 nM of one-site supercoiled circular p1 plasmid DNA (5). Reactions were performed at 37°C in reaction buffer (see above) supplemented with 4 mM ATP, 20 mM phosphocreatine and 5 U/ml creatine kinase. The reactions were initiated by adding 200 nM H.CglI and 200 nM R.CglI. Reactions were stopped by adding 2× Loading Solution (75 mM EDTA, 0.6% (w/v) orange G, 50% (v/v) glycerol, 0.5% (w/v) SDS, pH 8.0 at 25°C) followed by incubation at 70°C for 15 min. The reaction products were separated by electrophoresis through 1.0% (w/v) agarose gels with 0.5 μg/ml ethidium bromide. The relative amount of dsDNA cleavage was quantified by gel densitometry using BioDoc-Analyse image analyzer (Biometa, Göttingen, Germany) and the images were analyzed with OptiQuant 3.0 software (Packard Instrument) to determine the volume of each band, taking background readings into account. The fraction of DNA substrate left in each sample was calculated as volume(DNA substrate left)/(volume(DNA substrate left) + volume(DNA product)) (5). DNA cleavage rates are quoted as the mean of three independent experiments by fitting supercoiled DNA depletion data to a single exponential decay Equations (3) and (4).

### Small angle X-ray scattering (SAXS) experiments

SAXS data were collected at P12 EMBL beam line at PETRA III storage ring of DESY synchrotron in Hamburg (Germany). The SAXS data collection parameters and processing for H.CglI sample were described in (6). Dummy atoms models were calculated with DAMMIN (version 5.3) and DAMMIF (r6669) (51) applying P1 or P2 symmetry, as the molecular weight of H.CglI in solution corresponds to dimer (6). Both variants, P1 and P2 were very similar. Ten runs of each program were compared and averaged with DAMAVER package (52).

The DEAD-Z1 model and the crystal structure of the C-terminal domain were used for the rigid body modelling against solution scattering data with SASREF program (version 6.0 r6669) (53). As the molecular weight of H.CglI in solution corresponds to dimer (6), model was calculated without applying symmetry and contact conditions were used in order to keep domains interconnected.

### Data analysis

Quantitative analysis of protein-DNA binding and DNA cleavage was performed using KyPlot version 2.0 (54). The $K_d$ values for oligonucleotide binding by the CglI proteins were calculated by using a quadratic equation:

$$C = \left( d + p + K_d - \left[ (d + p + K_d)^2 - 4 \cdot d \cdot p \right]^{0.5} \right) / 2(1)$$

where $C$ is the formed protein–DNA complex concentration (nM) at each protein concentration $p$ (nM), $d$ is the DNA concentration in binding mixture, $K_d$ is the dissociation constant of protein DNA complex (nM).

The $K_d$ values for Mant-ATP binding by the CglI proteins were calculated by using a hyperbolic equation:

$$A = [A_0] + [A_1] \cdot p / (K_d + p) \quad (2)$$

where $A$ is the observed anisotropy values at each protein concentration $p$ (μM), $A_0$ is anisotropy observed with free
Mant-ATP, \( A_1 \) is the free amplitude, \( K_3 \) is the dissociation constant of protein–DNA complex (\( \mu M \)).

Rate constants \( (k_1) \) for DNA cleavage were calculated using one of the following equations:

\[
S = [A_1] \cdot \exp(-k_1 \cdot t)
\]

\[
S = [A_1] \cdot \exp(-k_1 \cdot t) + [A_2]
\]

where \( S \) is the amount of substrate left (nM), \( A_1 \) is the amplitude of the first phase, \( A_2 \) is the free amplitude, \( k_1 \) is the rate of the first phase, \( t \) is the reaction time in seconds.

RESULTS AND DISCUSSION

Crystallography trials of H.CglI and its domains

In order to understand the helicase-like mechanism of the H-subunit in more detail, we first tried to crystallise the full-length protein (Materials and Methods). A recombinant expression clone of H.CglI was produced and the subunit purified to >90% homogeneity (Supplementary Figure S1A). The CD spectrum suggested a folded protein (Supplementary Figure S1B) and assays demonstrating functional activity are described below. However, despite trying a wide range of conditions, we could not obtain crystals suitable for structure determination. One method to overcome the problem of inefficient crystallization is targeted mutagenesis of surface patches containing residues with large flexible side chains and their replacement with smaller amino acids (alanine). This can reduce surface entropy which in turn could help with formation of the crystal lattice (55). Mutants of the H.CglI were generated according to predictions using the SERp server (56): E22A + K23A (further EK mutant), Q116A + K117A (further QK mutant) and E361A + E362A (further EE mutant). The production and analysis of these mutants is described in the Supplementary Information and Supplementary Figures S1 and S2. Unfortunately, these mutant proteins also gave crystals unsuitable for structure determination. We nonetheless note that the predicted positions of the surface residues that were mutated are in agreement with the DEAD-Z1 model that is described below.

As an alternative approach, we reasoned that the individual domains of the H-subunit, or combinations thereof, might crystallize more efficiently. Recombinant expression clones for the isolated DEAD, Z1, C-terminal domains and the DEAD-Z1 and Z1-C-domain-fusions (6) were used to purify the proteins to >90% homogeneity (Supplementary Figure S1A) (Materials and Methods). The CD spectra were consistent with folded proteins (Supplementary Figure S1B), and assays demonstrating functional activity are described below. The DEAD, Z1, DEAD-Z1 and Z1-C proteins were refractory to crystallization. Mutants of the Z1 domain were generated according to predictions using the SERp server (55): Q295A+E296A (further QE mutant), E361A + E362A (further EE mutant) and K396A (further K mutant). These were also unsuccessful for structure determination. However, crystals of the C-terminal domain were generated, and a structure was solved, which is described in the next section.

Crystal structure of the C-terminal domain

Native and Se-methionine labelled C-terminal domains (residues 464–632) were crystallized by the sitting drop vapour diffusion method and the structure was determined by MAD (Materials and Methods). In the crystal structure 14 residues at the N-terminus and 9 residues at the C-terminus of the domain are disordered. The asymmetric unit of the C-terminal domain crystal contains one protein chain (Figure 2A, Supplementary Figure S3A). Analysis of the protein interfaces in the crystal with the PISA server (34) has not revealed any specific interactions that could result in the formation of stable quaternary structures indicating that the C-terminal domain is a monomer. This agrees with the gel filtration data (6). The C-terminal domain is folded into a 5-stranded mixed \( \beta \) sheet (\( \beta C3, \beta C5, \beta C6, \beta C7, \beta C1 \)) flanked by four \( \alpha \) helices, three helices from one side (\( \alpha C1, \alpha C2, \alpha C3 \)) and one helix (\( \alpha C4 \)) from the other side (Figure 2A, Supplementary Figure S3A). The domain also contains additional two-stranded antiparallel \( \beta \) sheet (\( \beta C2 and \beta C3 \)) and a helix \( \alpha C5 \). Structure comparison using DALI server (57) revealed that the C-terminal domain has a unique fold, which is also consistent with its lack of primary sequence similarity to other proteins (Supplementary Figure S3B). In isolation, the C-terminal domain structure does not provide an immediate answer to the question of its role.

Analysis of the C-terminal domain surface potential showed a patch of positively-charged residues located...
mostly on the surface of helices αC1, αC2 and αC3, namely K492, K499, K500, R517, K539 (Figure 2). This feature is characteristic of a DNA-interacting surface. However, there is no data indicating whether this domain is involved in DNA binding by H.CglI. Further biochemical analysis of the C-terminal domain alongside the other H-subunit domains is presented below.

The Z1 domain is characteristic of the second RecA-like domain (C-core) of an SF2 helicase

We reasoned that computational analysis and molecular modelling could be applied to H.CglI to provide information about the domain functions. In addition to the C-terminal domain structure determined above, the H-subunit amino acid sequence has two identifiable domains: (i) a RecA-like domain having Walker A and B motifs characteristic of ATP-dependent DNA/RNA helicases and other helicase-like enzymes (12); and (ii) a so-called Z1 domain (Pfam: PF10593). The Z1 domain having unknown structure and function was initially identified as a C-terminal domain of a SF2 helicase (17). However, at the time Z1 was predicted to have an α+β architecture, which is distinct from the typical α/β architecture of a SF2 C-core (RecA-like) domain. The H.CglI sequence was further analysed computationally, with particular focus on the uncharacterized Z1 domain. Iterative sequence searches dealing with multidomain proteins are prone to the overextension error (58). This type of error often occurs when one of the domains is very popular in nature (in this case DEAD domain), but the other domain (Z1) is uncommon. We therefore used a multi-step approach to provide a confident assignment of the Z1 fold.

Initially, H.CglI homologs were collected from UniprotKB using as a query only the H.CglI region encompassing Z1 and C-terminal domains. This was done in order to avoid inclusion of all the other P-loop ATPases that lack the Z1 domain. Remarkably, nearly all of the detected homologs of the H.CglI Z1 domain had the P-loop ATPase domain featuring Walker A and B motifs preceding the Z1 domain. This observation suggested that the Z1 domain might correspond to the universally present C-core domain of SF2 helicases. To test this hypothesis we used the alignment of H.CglI homologs (see Materials and Methods) as an input to HHpred (42) to search for related proteins in the PDB. The results revealed high confidence matches (HHpred probability >95%) to various DNA/RNA helicases and translocases. As expected, the H.CglI N-terminal region matched the N-core DEAD-like fold containing ATP-binding signature (Walker A and B motifs). Notably, the Z1 domain region confidently matched a C-core RecA-like domain indicating that H.CglI has two RecA-like core domains typical of the SF2 helicases, fused with the novel C-terminal domain.

The assignment of Z1 as the RecA-like C-core domain was further scrutinized by using the Z1 domain alone. The Pfam alignment for the Z1 domain sequences (filtered at 75% sequence identity) was downloaded and subsequently used as an input for the HHpred search against PDB. Again, the top matches were all against the C-core RecA-like domains of various helicases/translocases, albeit with slightly lower confidence (the highest HHpred probability around 90%). We therefore suggest that each H-subunit contains a standalone helicase-like motor comprising dual RecA folds.

A structural model of the DEAD-Z1 helicase-like motor of H.CglI

Taken together, the in silico analyses above show that the uncharacterized Z1 domain most likely represents a variant of the RecA-like C-core domain present in other SF2 helicases/translocases. Based on the detected homology with SF2 helicases/translocases, we constructed a structural model for both the N-core (DEAD) and the C-core (Z1) RecA-like domains.

To help understand how consistent the conserved amino acid sequence motifs are located within the putative RecA-like folds, we further sought to re-analyse the primary sequence alignment of the H-subunits to determine amino acid motifs that might be characteristic of this sub-family
of helicase-like proteins. The full-length H-subunit alignment used above was analysed for consensus sequences using WebLogo (59). We could clearly identify six characteristic helicase motifs (Figure 3A, Supplementary Figure S5): Motif I/Walker A, Motif Ia, Motif II/Walker B and Motif III within the DEAD domain, and, Motif V and Motif VI in the Z1 domain. The substitution of a glycine for a glutamine marks Motif I as being an uncommon Walker A variant amongst the SF1 and SF2 helicases. Another unique feature is motif VI, where instead of the typical QxxR signature, a shorter QxxR spacing is found, implying that the side chain of R408 is somewhat differently oriented compared to the ‘arginine-finger’ in a typical motif VI. Nonetheless, the presence of helicase Motif V and VI in the Z1 domain reemphasizes the potential importance of this domain to helicase-like motor activity.

A Y-motif characterised in other RE helicase-like subunits (60) could be also identified at the interface between the DEAD and Z1 domains. Additionally, we identified a Z1-specific motif, which we called H/H, which to our knowledge has not been observed in other helicases or helicase-like motors so far (Figure 3A). It also coincides with the beginning of an α-helical insertion/subdomain that seems unique to H.CglI and its close homologs. We note that the H/H motif is in a structurally-equivalent position to motif IV of SF2 helicases, which is most commonly involved in interaction with DNA/RNA. Motif IV does not have a common consensus sequence, but the H/H motif appears unique. We note that Motif IV sequences have been difficult to identify in Type I and III restriction enzymes (61). Nonetheless, we would predict a role for the H/H motif in DNA binding during translocation.

Clear motifs in the C-terminal domain could not be confidently discriminated, although solitary conserved residues were observed (e.g. N493, R562, W609, P622 and G632) (Supplementary Figure S5).

Overall, the common helicase motifs in the structural model of H.CglI are in the expected locations in 3-dimensions (Figure 3B). For comparison, the location of the SF2 helicase motifs in Rad54 is shown in Supplementary Figure S6. The N-core of CglI has a typical RecA-like structure with both Walker A (Motif I) and Walker B (Motif II, D158, E159 and D161) present at the ATP binding pocket. One of the specific features of the N-core of H.CglI and its homologs is that the Walker B (DEAD) motif is followed by an extended structurally-conserved loop (residues 163–172). The first half of this loop shows some sequence conservation (Supplementary Figure S5). Based on the proposed approximate location of the DNA, the more sequence divergent second half of Loop 163–172 has the potential to interact with the translocated DNA. A corresponding loop in an archaeal SW12/SNF2 protein (SsoRad54 dsDNA translocase), referred to as the ‘switch region’ or Motif IIa, has been shown to be involved in regulation of DNA-binding dependent ATPase activity (49). Corresponding regions in some other helicases, including the Type ISP REs, have also been implicated in the DNA-binding and/or ATPase activity (62–64).

The C-core has some unique features. One of these is a conserved α-helical insertion into the RecA-like fold. This insertion includes the two histidine residues (H282 and H290) of the conserved H/H motif identified above. Although some helicases/translocases do have domain/subdomain insertions in the corresponding region, the helical insertion seems to be unique to Z1 homologues. Furthermore, this insertion might be the reason why Z1 has not been documented previously as the C-core RecA-like domain. Although the modelled structure of the insertion has a low confidence, its relative position suggests that this structural motif might also be involved in binding the translocated DNA.

 Mutation of the helicase motifs in the Z1 domain abolish ATPase and nuclease activities

To begin to test our model of DEAD-Z1 as a dual RecA helicase-like motor, we expressed and purified to >90% homogeneity the following H.CglI mutants as full-length H-subunits (Supplementary Figure S1A): helicase motif mutants R379A (Motif V), Q405A (Motif VI) and R408A (Motif VI); H/H motif mutants H282A, H290A and H282A + H290A; and, Loop 163–172 mutants L1 (V169A), L3 (L165A + K168A + V169A) and ΔL (a deletion of the residues 165–172). According to size exclusion gel chromatography, all the H-subunits variants except L3 were dimeric and formed heterotetrameric R2H2 complexes with the wild type R-subunit (Supplementary Table S2). CD spectra of the Loop 163–172 variants showed that L3 was aggregated, so this mutant was not analysed further (Supplementary Figure S1B).

To evaluate characteristic enzyme activities of the wild type and mutant CglI proteins, we measured the following properties (Materials and Methods): (i) ATP-binding, as evaluated by a fluorescence anisotropy assay using Mant-ATP (65); (ii) ATP hydrolysis activity, measured from the phosphate release rates using a malachite green assay (6); (iii) DNA cleavage activity, measured by a gel-based assay using a one-site plasmid DNA which is an efficient substrate for the wild type enzyme; and (iv) DNA binding using an EMSA with a 25 bp non-cognate oligoduplex. Representative ATP binding, ATPase and DNA cleavage data is shown in Figure 4 for the loop and H/H motif mutants, while raw and fitted data for all the mutants are in Supplementary Figures S7–S10 and Tables S3–S6.

All of the variants were able to bind Mant-ATP with a similar affinity, within experimental error (1.9–5.0 μM), as the wild type H.CglI (3.2 μM, Figure 4 and Supplementary Table S6). The $K_d$ values for ATP binding by other DEAD-box proteins can range between ~50 nM and >6 mM (66–68). The CglI values are not atypical and SF2 helicase RecG has a similar $K_d$ for Mant-ATP of ~4 μM (69), while NS3 has a $K_d$ for ATP of ~2 μM (70). In the case of the R379A and R408A mutants, the amplitude of the anisotropy change was lower than that of the wild type protein, but they could nonetheless still bind to the Mant-ATP (Supplementary Figure S11). However, all of the variants had significantly decreased or unmeasurably-low ATPase activity and DNA cleavage activity close to, or at, background rates (Figure 4 and Supplementary Figures S8–S9 and Tables S4–S5). This suggested that although ATP can bind in all cases, the coupling between ATP turnover and subsequent DNA cleavage has been broken. Similarly, mu-
tations in motif VI of RecD2 only slightly reduced affinity for nucleotide or RNA binding, but decreased ATPase turnover by 10-fold, which is also observed in other DEAD-box helicases (71). Motif VI most likely plays a direct role in ATP catalysis, typical of its role in other helicases and helicase-like enzymes. Structures equivalent to Motif V and Loop 163–172 are implicated in DNA interactions. However, in our EMSAs, little difference was observed between the wild type H.CglI and the loop, helicase or H mutants. Structure equivalent to Motif V and ATP catalysis, typical of its role in other helicases and helicase-like enzymes. The DNAbindingdatadoesnotsupportarolefortheH motif in DNA interactions (Supplementary Figures S7 and S8) which is consistent with the results above. Again, in the absence of ATP, an H motif mutant of the Z1 domain and Z1-C fusion had an apparent $K_d$ values that were lower than the full-length protein ($K_d = 3.2$ or 3.7 $\mu$M, respectively). This provides compelling evidence in support of the model in Figure 3B for the DEAD-Z1 domains being the complete helicase motor.

Titration of the wild type H-subunit against a non-cognate oligoduplex produced a band-shift on a native polyacrylamide gel, consistent with DNA binding (Figure 5B). In addition to a discrete shifted band, a ‘smearing’ of the free DNA was observed which most likely represents more weakly associated complexes that dissociate during the electrophoresis process. DNA binding activity was observed with the isolated Z1 domain and C-terminal domains, and with the DEAD-Z1 and Z1-C fusions, but not with the DEAD domain alone (Supplementary Figures S7 and S8). The isolated Z1 domain and Z1-C fusion had an apparent $K_d$ values that were lower than the full-length protein ($K_d = 46$ nM) or the DEAD-Z1 fusion ($K_d = 844$ nM) (Supplementary Table S3), which may indicate that the Z1 domain is autoinhibited by the DEAD domain. The data suggests that Z1-C provides an important DNA contact surface. The lack of binding of DNA by the isolated DEAD domain is perhaps surprising given our prediction of a role in DNA transport and the presence of Loop 163–172. However, it may be that Loop 163–172 only makes transient DNA contacts as part of a mecanochemical cycle that cannot be mimicked by EMSAs in the absence of ATP. An H/H mutant of the Z1 domain had an apparent $K_d$ similar to wild type ($K_d = 99$ nM) (Supplementary Figures S7 and S8 and Table S3) which is in agreement with the results above. Again, in the absence of ATP we cannot fully rule out a role in DNA interaction during translocation.
Figure 5. Functional analysis of the domains of H.CglI. (A) Mant-ATP binding by the full-length H.CglI and its isolated domains was studied by fluorescence anisotropy measurements. Data points were fitted to a hyperbolic function. (B) DNA binding by the full-length H.CglI and its domains measured using an EMSA. The reactions contained 1 nM of the 32P-labelled oligoduplex and the proteins indicated at increasing concentrations (nM): 0, 50, 100, 200, 500, 1000, 2000 and 5000 (shown as wide black triangles). Images of separate native polyacrylamide gels are shown, obtained by Fujifilm FLA-5100 fluorescent image analyzer and the OptiQuant 3.0 software. Filled and open triangles correspond to the mobility of the free DNA and a protein–DNA complex, respectively. (C) Cleavage of the 21-site pBR322 plasmid DNA by the mixtures of the R.CglI and H.CglI or its domains, as indicated, for 1 h at 37 °C. M – marker (DNA fragments of known length in bp), K – control (pBR322 plasmid DNA), SC – supercoiled circular DNA, OC – open circular DNA (‘nicked’), FLL – full-length linear DNA, CP – cleavage products at multiple sites. (D) Rate of cleavage of the first strand (nicking) of one-site supercoiled DNA by the wt R2H2.CglI and R2HΔH.CglI (lacking a single C-terminal domain within H.CglI) complexes. Data points are fitted to a single exponential function. The fitted values for panels A, B and C are presented in Supplementary Tables S3, S4 and S6. All data points are the average of 3 or more reactions, with the error bars ± SD.

Given the availability of the C-terminal domain structure and its DNA binding activity, we also mutated residues on the basic surface patch (Figure 2), in both the isolated domain and the full-length H-subunit: R517E, K539E and 3M (K492E+K499E+K500E). We tested DNA binding, ATPase and nuclease activities (Supplementary Information, Supplementary Figures S7-S10, Supplementary Tables S3-S5). The results support a role for the C-terminal domain basic surface patch in binding to DNA and in the coupling DNA binding to ATP hydrolysis and long-range communication necessary for DNA cleavage (Supplementary Information).

The importance of H-subunit domain connectivity and complete heterotetrameric assembly to the activity of CglI

Since the isolated domains and domain fusions of H.CglI have biochemical activity, we reasoned that it may be possible to reconstitute full RE activity using different combinations of the domains/domain fusions along with the full-length R.CglI. We tested this by mixing different protein combinations at 1:1 stoichiometry, and then following the cleavage of pBR322 (contains 21 CglI target) at a reaction endpoint (37°C for 1 h), conditions under which the wild type proteins cut the DNA multiple times (Figure 5C). Although background levels of DNA cleavage were observed with some combinations, only the full-length H-subunit produced efficient DNA cleavage activity with the R-subunit. A similar conclusion could be drawn from measurements of ATP hydrolysis using combinations of the domains (Supplementary Table S5). This may be the consequence of inter-domain motions during ATP hydrolysis that cause disassembly of the complex unless the domains are connected as a single polypeptide. We note however that the DEAD-Z1 fusion was still capable of ATP hydrolysis, albeit 20-fold reduced compared to the full complex.

This data suggests that the DEAD, Z1 and C-terminal domains are all required for full RE activity. A key mechanistic question is why CglI is active as a tetramer with two H-subunits. Given our model and data, there are two separate helicase motors (Figure 1B). In the cleavage model, an R2H2 complex must translocate along DNA to commu-
nicate with a second enzyme at a distant target site. However, it isn’t immediately obvious why two motors would be needed. To begin to approach this problem, we isolated an R2H2 CglII complex without one C-terminal domain (R2HΔ1H.CglII) (Supplementary Information, Supplementary Figure S4). This complex assembles into the expected heterotetrametric complex (Supplementary Table S2). Surprisingly, this complex had a >5-fold reduction in ATPase activity (Supplementary Table S5) and a corresponding ~60-fold reduction in the rate of cleavage of a one-site plasmid DNA (Figure 5D, Supplementary Table S4). Despite not being part of the dimerization interface (6), it appears that the C-terminal domains of both H-subunits are necessary for efficient ATPase and DNA cleavage activities.

Small angle X-ray scattering analysis of H.CglII

Based on gel filtration analysis (Supplementary Table S2) and the $M_w$ calculated from SAXS data (76), H.CglII is a dimer in solution with the dimer interface being the RecA-like Z1 domains (6). In the absence of diffracting crystals, we used small angle X-ray scattering to provide further structural information on the overall shape of the H.CglII dimer that could help refine the model in Figure 1B and which could help to resolve whether both motors are cooperating in DNA translocation.

The ab initio estimation of an overall shape of the particle filled by dummy atoms is presented in Figure 6A. According to SAXS data H.CglII dimer in solution has an elongated form.

We also performed rigid body modelling of the H.CglII dimer using the C-terminal domain crystal structure and the DEAD-Z1 model as separate domains comprising full length H.CglII (Figure 6A). The program used (SASREF 6.0 (53)) allows us to use additional information about contacts between domains. Based on dimerization via the Z1 domains, we placed two copies of the DEAD-Z1 model into the centre and connected a few distal amino acids Cx atoms of both Z1 domains (S253–S253 ~8 Å and N246–A267 ~9 Å). The limits are rather loose, reflecting our uncertainty regarding the exact dimerization interface. According to our ab initio shape determination the monomers of H.CglII are also elongated. Based on the proposed orientation of DNA bound by the DEAD-Z1 model we have connected (by very loose restraints), the positively charged surface of the C-terminal domain to residues located in the groove between the DEAD and Z1 domains (N211–D488 ~8 Å, D186–V494 ~7 Å, E310–R547, ~41 Å and D40–Q593 ~10 Å).

These constraints were used in order to keep the presumed DNA-contacting interfaces of DEAD-Z1 model (Figure 3B) and C-terminal domain (Figure 2). As the unmodelled linker between Z1 and C-terminal domains comprises 30 amino acids (residues 448–477), no connectivity between C-terminus of the Z1 domain and N-terminus of the C-terminal domain was used as modelling restraint. Initial position of C-terminal domain was restrained at the distal end of the DEAD domains in the initial orientation of domains based on the ab initio shape obtained from the averaged dummy atom models and assuming dimerization via Z1-domains. The rigid body model obtained could be overlaid with the dummy atoms model (Figure 6A). The comparison of the scattering curve calculated from the rigid body model with the experimental scattering curve is shown in Figure 6B. The comparison shows that the model is smaller than the averaged shape of the particle in solution, as the I(s) values of the model are lower than experimental data. One reason for this could be partial aggregation of the protein sample.

CONCLUSIONS

According to homology modelling and identification of conserved sequence motifs, both the DEAD and Z1 domains of CgII are RecA-like domains that together form an interface where ATP binds and is contacted by key catalytic
motifs such as Walker A, Walker B and Motif VI. Biochemical assays using wild type and mutant proteins are consistent with the model (Figure 3B). The Z1 domain is more distantly related RecA-like fold, which may be why it has not been identified as a helicase family domain up to this point (17). We propose that H.CglI is a prototype of a new SF2 helicase sub-family that is both structurally- and functionally-related to, but distinct from, other RE helicase-like motors (Figure 7). The DEAD-Z1 helicase-like motor has several characteristic features which are necessary for ATP and DNA hydrolysis; (i) an unusual Walker A motif sequence compared to other SF1 and SF2 helicases. Some Type ISP-like REs also have an unusual RFGKT Walker A sequence (77), but we note that in the wide P-loop family the first glycine is not conserved well in the Walker A motifs; (ii) an unusual register of the ‘arginine finger’ in the Motif VI compared to other SF1 and SF2 helicases. These Type ISP-like REs also have an unusual RFGKT Walker A sequence (77), but we note that in the wide P-loop family the first glycine is not conserved well in the Walker A motifs; (ii) an unusual register of the ‘arginine finger’ in the Motif VI compared to other SF1 and SF2 helicases; (iii) a loop following Motif II in the DEAD domain. A region that is similarly located in Type ISP REs, but which follows Motif III (Figure 7), may also play a role as a ‘molecular paddle’ during dsDNA translocation (78); and (iv), a previously undescribed H/H motif and structural insertion in the Z1 domain. This motif couples ATP hydrolysis to DNA motion/cleavage but its exact role remains to be confirmed. Although the modelling predicted a role in DNA interaction (equivalent to Motif IV), mutation of the H/H motif did not affect DNA binding activity. Despite the Type I, Type ISP and CglI proteins all undertaking similar dsDNA translocate functions, because of the differences between the proteins at the sequence and structural level, it remains difficult to predict the function of helicase-like proteins without first undertaking biochemical assays.

Although the DEAD-Z1 motor has basic ATP and DNA binding activities, and some residual ATPase activity, the complete H-subunit including the C-terminal domain is required for full activity. We were able to solve a structural model for the C-terminal domain of H.CglI, showing that it possesses a unique fold with a basic amino acid interface. We could demonstrate that the Z1 and C-terminal domains are responsible for interaction of H.CglI with DNA, but we cannot rule out that other domains or motifs are key to the dsDNA translocation activity. Interestingly, both full-length H-subunits within the R2H2.CglI complex were required for catalytic activity of the complex. We can now confirm that the complex has two complete helicase-like DEAD-Z1 motors and that the H-subunit dimer forms an elongated shape. We suggest a possible structure for the R2H2.CglI complex in Supplementary Figure S12. Whether both motors cooperate in an individual dsDNA translocation event remains to be determined. Future structural work could exploit recent advances in cryo-EM to obtain a structure for the complete R2H2 complex bound to DNA. In the meantime, our improved understanding of the H-subunit domains and motifs will help in more advanced biophysical assays that we are developing to follow the dsDNA translocation activity.

**DATA AVAILABILITY**

Atomic coordinates and structure factors of the H.CglI C-terminal domain crystal structure has been deposited with the Protein Data Bank under accession number PDB ID 6F1S.
SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Authors acknowledge Audra Ruksenaite for mass spectrometry, Gleb Burenkov and Alexey Kikhyne for assistance with data collection at EMBL P14 and P12 beamline at the PETRA III storage ring, DESY Hamburg. The access to synchrotron has received funding from the European Community’s Seventh Framework Programme (FP7/2007–2013) under BioStructX_7869.

FUNDING

Research Council of Lithuania [MIP-56/2015 to M.Z.]; NIH [1R01GM111507-01 to M.D.S.]. Funding for open access charge: Research Council of Lithuania.

Conflict of interest statement. None declared.

REFERENCES


