 doi:10.20944/preprints202105.0161.v1

Article

Mini-intein structures from extremophiles suggest a strategy for finding novel robust inteins

Mimmu K. Hiltunen[1], Hannes M. Beyer[1,†] and Hideo Iwai [1,*]

[1] Institute of Biotechnology, HiLiFE, University of Helsinki, P.O. Box 56, 00014 University of Helsinki, Finland

[†] Current affiliation: Institute of Synthetic Biology and CEPLAS, University of Düsseldorf, 40225 Düsseldorf, Germany

* Correspondence: hideo.iwai@helsinki.fi

Abstract: Inteins are prevalent among extremophiles. Mini-inteins with robust splicing properties are of particular interest for biotechnological applications due to their small size. However, biochemical and structural characterization has still been limited to a small number of inteins, and only a few inteins serve as widely used tools in protein engineering approaches. We determined the crystal structure of a naturally-occurring Pol-II mini-intein from *Pyrococcus horikoshii* and compared it with two other natural mini-inteins from *Pyrococcus horikoshii*. Despite the similar sizes, the comparison revealed distinct differences in insertions and deletions, implying specific evolutionary pathways from distinct ancestral origins. Our studies suggest that sporadically distributed mini-inteins might be more promising for further protein engineering applications than the highly conserved mini-inteins. Structural investigations of more inteins could guide the shortest path to finding novel robust mini-inteins suitable for protein engineering purposes.

Keywords: protein splicing; intein; crystal structure; hyperthermophile; protein engineering

1. Introduction

Self-splicing protein intron (inteins) are genetic elements that are translated with their host proteins [1,2]. After translation, inteins catalyze their own excision and re-ligation of flanking protein regions (called exteins) with a peptide bond, resulting in active mature host proteins [1,3]. Inteins are considered selfish genetic elements found within coding DNA regions because their removal generally does not affect the fitness of host organisms [3]. However, for some inteins specific regulatory functions of inteins have been proposed [4,5].

Inteins are prevalent among extremophiles, such as thermophiles and extreme halophiles, suggesting that they might play important roles under extreme conditions or have ancient origins that predate the separation of prokaryotes and eukaryotes [4,5,6,7,8]. The closely related thermophiles *Pyrococcus abyssi* (*Pab*) and *Pyrococcus horikoshii* (*Pho*), for example, both contain 14 inteins in their genome, while the halophile *Haloquadratum walsbyi* contains 15 inteins [6]. However, the distribution and size of inteins are not highly conserved even among thermophilic archaea of the same genus. Presumably, horizontal gene transfers (HGT) that occurred during evolution, counterbalanced by degeneration events of the nested homing endonuclease domains (HENs) contribute to this variation [3]. Due to these evolutionary events, the structure and protein-splicing activities of inteins might represent the evolutionary history of each intein [9].

Studies of the obscure biological roles of inteins and their enzymatic mechanisms catalyzing protein splicing reactions have opened a new horizon in protein chemistry. Utilizing the protein splicing activities of inteins bears a repertoire of potential applications in several areas, including *in vivo* protein engineering, protein purification, and modification. Indeed, intein-mediated chemical reactions have increasingly been incorporated as practical tools in the fields of protein engineering, synthetic biology, and biotechnology [10,11]. For example, inteins from extremely halophilic archaea have been demonstrated

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2021 doi:10.20944/preprints202105.0161.v1

to control protein-splicing reaction with salt concentrations, which has enabled the engineering of a salt-inducible self-cleaving tag for protein purification [12]. Thus, inteins from extremophiles might have great potential for the developments of unique biotechnological tools.

Inteins often contain a nested active or inactive HEN, which presumably plays an essential role in HGT of intein genes [3,6,13]. There is a class of inteins lacking the HEN domain, so-called mini-inteins. Due to their reduced complexity, naturally-occurring mini-inteins lacking HEN domains have been of special interest for protein engineering to develop biotechnological applications [14,15]. Although the protein-splicing HINT (Hedgehog/INTein) and DNA-processing HEN domains likely function independently of each other [1], attempts to engineer mini-inteins by removing the HEN have revealed a more complicated relationship between the two domains [9, 15,16,17]. Some engineered mini-inteins retain their splicing activity, while others lose it completely [9,14,15,16].

Currently, inteins have not been systematically selected and tested for the robustness of their protein splicing activity because their splicing activties are not predictable without experimental assesments. Thus, some strategies for selecting promising inteins from sequence databases would be highly desirable for protein engineering.

As the first step towards a rational approach to identify robust inteins for protein engineering, we turned our attention to the 14 inteins in *P. horikoshii*. Three out of these inteins can be classified as mini-inteins (≤200 residues). These are the *Pho*RadA, *Pho*CDC21-1, and *Pho*Pol-II inteins, consisting of 172, 170, and 166 residues, respectively. The structural and biochemical characterization of *Pho*RadA and *Pho*CDC21-1 inteins have been previously reported [17,18]. In this work, we determined the 1.48-Å resolution structure of the *Pho*Pol-II intein, the smallest inteins found in *P. horikoshii*, and compared the selected inteins of the *Pyrococcus* genus with other naturally occurring mini-inteins.

2. Materials and Methods

2.1. Cloning and Production of Pho*Pol-II intein*

The gene encoding the *Pho*Pol-II intein with Cys1Ala (C1A) mutation to inhibit self-cleavages during the purification was amplified from pSKDuet23 [19] using the two oligonucleotides HK941: 5′-AGGATCCGGTAATGCCTTCCCGGGAGATACAAG and HK942: 5′-TGAAAGCTTACTGATGCGTCACAATATTTTC. The PCR product was cloned between *Bam*HI/*Kpn*I of pHYRSF53 (Addgene #64696), resulting in plasmid pCARSF55D [19]. The plasmid pCARSF55D encodes the N-terminal H_6-SUMO domain (yeast SMT3) fused with the *Pho*Pol-II intein with C1A mutation. The C-terminal residue was kept as the original glutamine (Gln), followed by the stop codon, resulting in no C-extein residue. The *Pho*Pol-II intein was produced in *E. coli* strain T7 Express (New England Biolabs) using plasmids pCARSF55D and pRARE [18]. The transformed cells were grown at 37 °C in 2-liter LB expression cultures supplemented with 25 μg mL^{-1} kanamycin and 5 μg mL^{-1} chloramphenicol. The cultures were induced with a final concentration of 1 mM isopropyl-β-D-thiogalactoside (IPTG) for 3 h when OD_{600} reached 0.6. The induced cells were harvested by centrifugation at 4700×g for 10 min, 4 °C and lyzed in 20 mL Buffer A (50 mM sodium phosphate, pH 8.0, 300 mM NaCl) using an EmulsiFlex-C3 homogenizer (Avestin Inc, Ottawa, Canada) at 15,000 psi for 10 min, 4 °C. Lysates were cleared by centrifugation at 38000 ×*g* for 60 min, 4°C. The *Pho*Pol-II intein (C1A) was purified using a 5 mL HisTrap HP column (GE Healthcare Life Sciences) as previously described, including the removal of the N-terminal H_6-SUMO fusion domain [19]. The protein was dialyzed against deionized water and concentrated for crystallization using Macrosep® Advance Centrifugal Devices 10K MWCO (PALL Corporation, New York, USA).

2.2. Cis-splicing of Pho*Pol-II intein*

For the cis-splicing test of *Pho*Pol-II intein, the gene of the active *Pho*Pol-II intein, including sequence encoding two residues of ”GN” and ”CD” at the N- and C-terminal splicing junction, respectively, was amplified from pCARSF55D using J603: 5′-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2021 doi:10.20944/preprints202105.0161.v1

AAGGATCCGGTAATTGCTTCCCGGGAGATACAA and J618: 5′-TAGGTACCATCG-CACTGATGCGTCACAATATTTTC. The expression vector for the *cis*-splicing precursor with two the B1 domain of *Staphylococcus* protein A (GB1) as the two exteins was created by cloning the PCR product between *Bam*HI/*Kpn*I of SKDuet16, resulting in pLKRDuet30. The *cis*-splicing precursor protein with two GB1 as the exteins was produced in *E. coli* strain T7 Express (New England Biolabs) using plasmids pLKRDuet30 and pRARE as described above. The precursor protein was purified using a 5 mL HisTrap HP column (GE Healthcare Life Sciences) and dialyzed against PBS. The purified precursor protein was incubated in the presence of 0.5 mM TCEP at either room temperature, 37 °C, or 50 °C. The samples were taken at 0, 1, and 3 hours, and overnight (ON), and were analyzed by SDS-PAGE on 16.5% acrylamide gels, and visualized using Coomassie Blue staining.

2.3. Crystallization of Pho*Pol-II inteins*

23.4 mg/ml of *Pho*Pol-II intein (C1A) was used for crystallization trials. Drops of 200 nl (100 nl concentrated protein and 100 nl of reservoir solution) were set up in 96-well MRC (Molecular Dimensions) crystallization plates using a Mosquito LCP® (TTP Labtech, UK). Diffracting crystals were obtained with the reservoir solution containing 100 mM MES pH 6, 15% (w/v) PEG 550 MME, and 30 mM zinc sulfate. 25% PEG MME 550 was added as a cryoprotectant for flash-freezing crystals in liquid nitrogen. The *Pho*Pol-II intein diffraction data were collected on beamline I03 at the Diamond Light Source with Eiger2 XE 16M detector (Oxfordshire, UK) and were subsequently indexed, integrated, and scaled to a 1.48-Å resolution using the program XDS [21].

2.4. Structure determination and refinement

The crystal structure of *Pho*Pol-II intein was solved by molecular replacement. The search model was modeled using SWISS-MODEL based on the primary structure [22]. The initial solution obtained from Phaser using the search model was used for auto-building by ARP/WARP [23]. The initial coordinate obtained from ARP/WARP was rebuilt with Coot, followed by rounds of refinement using the Phenix software [24,25]. The polypeptide chain of *Pho*Pol-II intein was easily traced into the electron density map without breaks for 168 residues, 166 of which belong to the intein. We used the comprehensive validation tool in the Phenix GUI for validating the quality of the final structure (Table 1) [25].

3. Results

3.1. Crystal Structure of PhoPol-II intein

DNA polymerase II large subunit (Uniprot: O57861) of *P. horikoshii* contains 166-residue intein (*Pho*Pol-II intein). Previously, Pol-II intein from *P. abssi* (*Pab*Pol-II intein) has been solved by NMR spectroscopy [26]. Whereas *Pab*Pol-II consists of 186 residues, *Pho*-Pol-II intein only has 166 residues, which makes the latter more attractive for protein engineering due to its smaller size. We solved the three-dimensional structure of *Pho*Pol-II intein at the 1.48-Å resolution by molecular replacement (Figure 1). The crystal structure revealed the typical HINT fold with the β-strand insertion commonly observed among inteins from thermophilic organisms (Figure 1) [27]. In line with the apparent thermophilic structure minimization [28], we could trace electron densities for all the 168 residues without detecting any flexible linker sequences which were present in the structure of the *Pab*Pol-II intein (Figure 3b) [26].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2021 doi:10.20944/preprints202105.0161.v1

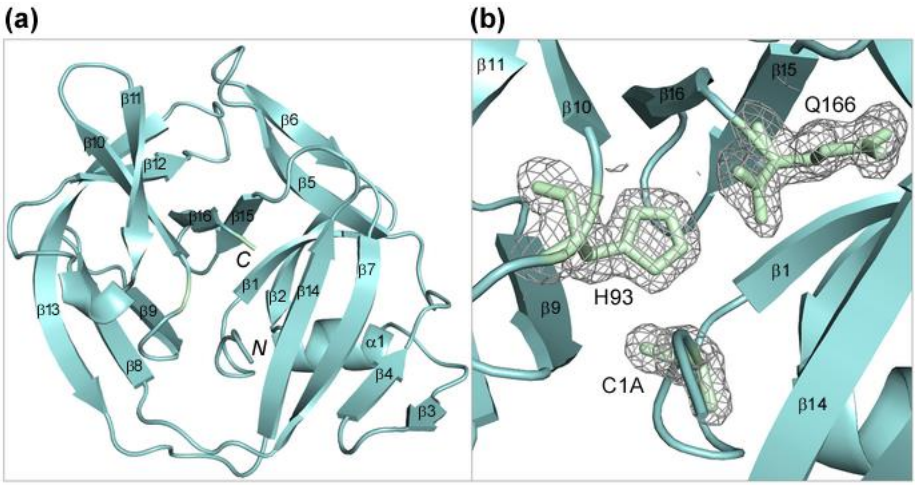

Figure 1. (**a**) A cartoon presentation of *Pho*Pol-II intein. (**b**) A close-up of the active site with the electron densities for Cys1Ala (C1A), His93 (H93), and Gln166 (Q166). *N* and *C* indicate the N- an C-termini, respectively.

Table 1. Data collection and structure refinement

Intein	*Pho*Pol-II intein (C1A)
PDB ID	7OEC
Data collection	DIAMOND I03
Space group	P 4_1 2_1 2
Cell dimensions	
a, b, c, Å	70.82, 70.82, 70.66
α, β, γ, °	90.00, 90.00, 90.00
Wavelength, Å	0.9763
Resolution, Å	70.89-1.42 (1.44-1.42)
Total reflections	776435 (76171)
Unique reflections	30575 (2978)
Completeness, %	99.87 (99.77)
I/σ	16.35 (4.23)
R_{meas}[a]	0.1939 (7.025)
$CC_{1/2}$[c]	0.998 (0.572)
Multiplicity	25.4 (25.5)
Refinement	
Molecules/*au*	1
Resolution, Å	35.41 - 1.48 (1.533 - 1.480)
Reflections (refinement /R_{free})	30534/1527
R_{work} / R_{free}[b]	0.1537/0.1873
Number of atoms	
Protein	1382
Water	76
Ligand	34
RMS deviations	
Bond length, Å	0.015
Bond angles, °	1.34
Ramachandran plot, %	
Most favored regions	97.55
Outliers	0.00
Average B-factors, Å^2	29.90
Protein	28.41

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2021 doi:10.20944/preprints202105.0161.v1

Water	37.48
Clash score	2.12
Molprobity score	0.97

Numbers in parentheses represent the highest resolution shell.
au, asymmetric unit
[a]$R_{meas}=\Sigma_h[n/(n-1)]^{1/2}\,\Sigma_i\,|I_i-\langle I\rangle|\,/\Sigma_h\Sigma_i\,I_i$, where Ii is the observed intensity of the ith measurement of reflection h, ⟨I⟩ is the average intensity of that reflection obtained from multiple observations, and n is the multiplicity of the reflection
[b]$R=\Sigma||F_o|-|F_c||\,/\Sigma|F_o|$, where F_o and F_c are the observed and calculated structure factors, respectively, calculated for all data. R_{free} was defined by Brünger [29].
[c]$CC_{1/2}$ was defined by Karplus et al. [30].

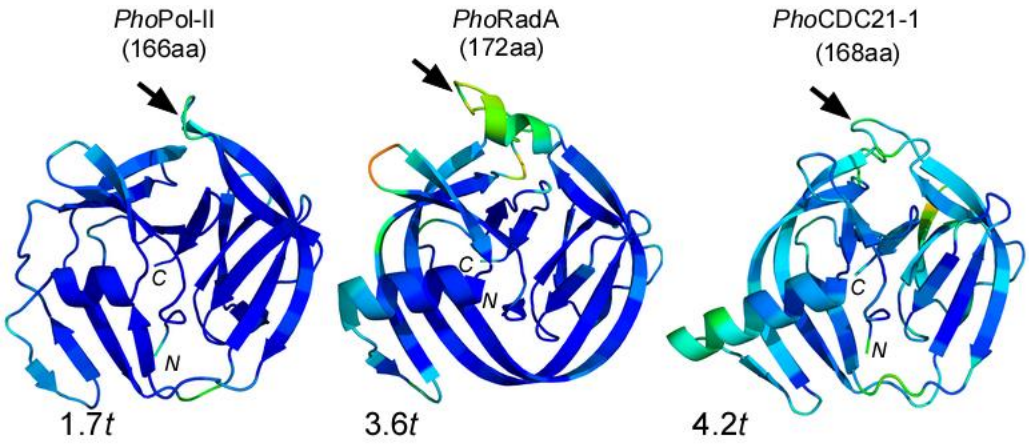

Figure 2. Cartoon representations of the structures of *Pho*Pol-II, *Pho*RadA, and *Pho*CDC21-1 inteins. The views are from the dorsal side [18]. Chain A from the coordinate (4e2t) was shown for *Pho*RadA intein [17]. *N* and *C* indicate N- and C-termini, respectively. The color codes represent the B-factor. The arrows indicate the insertion sites commonly observed for the homing endonuclease (HEN) domain.

3.2. Comparison with other mini-inteins from P. horikoshii

Although the three mini-inteins in *P. horikoshii* have similar sizes between 166-172 residues, their three-dimensional structures show distinct differences (Figure 2). All three structures have the extended β–strand insertion (extβ) commonly observed among thermophilic inteins [27]. The helical lengths preceding extβ vary among the three structures with 1.7, 3.6, and 4.2 helical turns for *Pho*Pol-II, *Pho*RadA, and *Pho*CDC21-1 inteins, respectively. *Pho*RadA intein has a 10-residue insertion at the highly conserved homing endonuclease domain (HEN) insertion site that could be removed without affecting the splicing activity [17]. In contrast, *Pho*Pol-II and *Pho*CDC21-1 inteins lack such insertions at the HEN insertion site (Figure 2). In fact, the relatively short helix of *Pho*Pol-II intein is rather reminiscent of a canonical mesophilic intein. Despite their similar sizes, the structural comparison suggests that the three inteins are unlikely to be evolved from the same ancestral intein.

3.3. Comparison with naturally-occurring mini-intein structures from other thermophiles

We subjected the coordinate of *Pho*Pol-II intein to the DALI protein structure comparison server in order to identify the closest three-dimensional structures. Not surprisingly, the server returned *Pab*Pol-II intein (2LCJ) as the closest structure with a Z-score of 22.7, covering 164 residues with 1.5 Å-RMSD for Cα atoms (Figure 3) [26]. *Pab*Pol-II intein shares 70% sequence identity with *Pho*Pol-II intein (Figure 3a). The largest difference between the two inteins lies in the flexible 19-residue sequence at the HEN insertion site for *Pho*Pol-II intein (Figure 3b). We assume that the insertion is a remnant of HEN

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2021 doi:10.20944/preprints202105.0161.v1

degradation which occurred during the evolution. The structures of CDC21-1 inteins from *P. horikoshii* and *P. abssi* have Z-scores of 22.0 and 22.2, respectively (Figure 2) [18]. These two inteins are closely related to the same insertion site in the CDC21 protein. However, the longer helices in CDC21-1 inteins are very different from Pol-II inteins. The observed evident variation in the helix length within the thermophilic insertion might be caused by differences in the evolutionary origins of the different inteins. The VMA mini-intein from *Thermoplasma volcanium* (*Tvo*VMA) has the same Z-score of 22.2. However, it resembles the *Pho*RadA intein, because both structures share a prominent extension of the helix [14]. The KlbA intein from *Methanococcus jannaschii* (*Mja*KlbA) is another naturally-occurring mini-intein and has a Z-score of 19.1 with *Pho*Pol-II intein [31]. All the three-dimensional structures mentioned show the typical HINT fold, including the additional β-strand insertion (extβ) preceded by a helix of variable length, even though the growth temperatures of their hosts vary drastically between 33-104°C (Table 2).

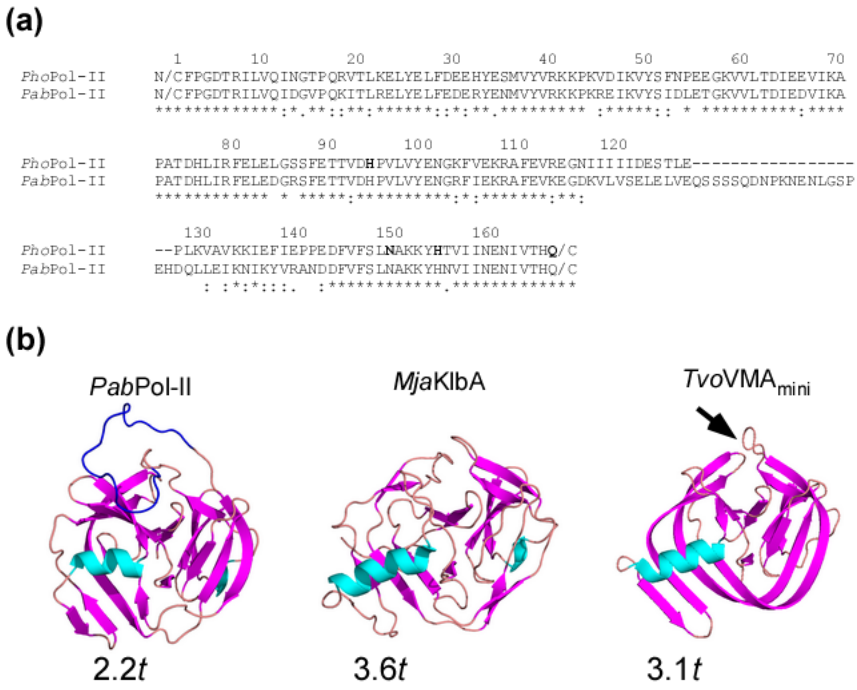

Figure 3. (a) A sequence alignment of Pol-II inteins from *P. horikoshii* and *P. abyssi*. **(b)** Cartoon models of the three-dimensional structures of mini-inteins from other thermophilic organisms: *Pab*Po-II intein (left), *Mja*KlbA intein (middle), and *Tvo*VMA intein (left). PDB coordinates of 2LCJ (*Pab*Po-II), 2JNQ (*Mja*KlbA), and 4O1S (*Tvo*VMA) were used. β-strands and helices are colored in magenta and cyan, respectively. The lengths of helices are indicated by the number of helical turns (3.6 residues per helical turn). The flexible loop of *Pab*Pol-II is colored in blue. The location of the loop in *Tvo*VMA intein, of which 21 residues were removed for the crystallization, is indicated an arrow. The number of helical turns is indicated below each structure.

3.4. Protein-splicing activity of Pho*Pol-II intein*

While the hyperthermophilic archaeon *Pyrococcus horikoshii* grows at a temperature between 88-104 °C (Table 2) [32], the *Pho*RadA intein is capable of efficiently catalyzing protein splicing at ambient temperatures *in vitro* and in *E.coli* cells [17]. This observation suggests that the protein splicing reaction is unlikely a rate limitation during the biosynthesis of active RadA protein. At ambient temperatures, the *Pho*CDC21-1 intein retains protein-splicing activity with foreign exteins to a lesser extent than *Pho*RadA intein in *E.coli*, but its splicing activity could be improved by increasing the temperature [19].

Conversely, *Pho*Pol-II intein was found inactive at ambient temperatures [19]. At a higher temperature and longer incubation, *Pho*Pol-II intein showed weaker protein-splicing activity as well as side reactions such as cleavages (Figure 4). Due to this delicate activity profile, it appears plausible that the *Pho*Pol-II intein might contribute to the physiological regulation of active DNA polymerase II production. In other words, the production

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 May 2021 doi:10.20944/preprints202105.0161.v1

of active DNA polymerase might be dependent on the protein splicing activity of *Pho*Pol-II intein, which could be a survival strategy of mini-inteins lacking the HEN domain to protect themselves from removal. Both *Pho* and *Pab* Pol-II inteins have an atypical C-terminal Gln residue instead of the canonical Asn responsible for cleavage of a branched intermediate during the protein-splicing reaction (Figure 1b and 3) [26,33]. The ^{1}H-^{15}N correlation peak for the C-terminal Gln was not visible for *Pab*Pol-II intein in HSQC spectra suggesting the slow conformational exchanges rather than a fixed conformation [26]. The sidechain of the C-terminal Gln166 in *Pho*Pol-II intein shows clear electron density (Figure 1b). We also introduced a stop codon after the last residue of Q166 so that any cleavage-related mechanism by the C-terminal Gln could not occur.

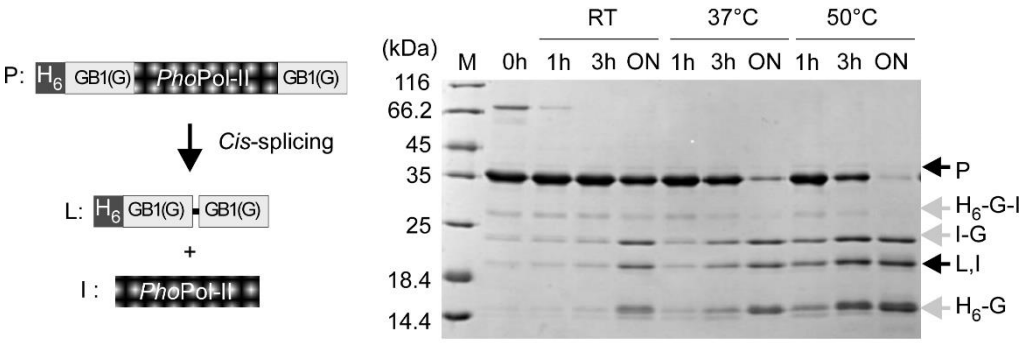

Figure 4. *In vitro* protein splicing in *cis* by *Pho*PolII-intein. The precursor protein containing the *Pho*PolII-intein (P) was purified and incubated at room temperature, 37 °C, and 50 °C.

Table 2. Summary of naturally occurring mini-inteins from thermophiles and their growth temperatures.

Organism	Growth Temp.	Gene	PDB
Pyrococcus horikoshii	88 -104 °C (98 °C) [32]	*radA*	2LQM, 4E2U, 4E2T
		cdc21	6RPQ
		pol-c	7OEC
Pyrococcus abyssi	68 to 102 °C (96 °C) [34]	*pol-c*	2LCJ
		cdc21	6RPP
Thermoplasma volcanium	33-67 °C (60°C) [35]	*vma*	4O1S
Methanococcus jannaschii	48-94°C (85°C) [36]	*klbA*	2JMZ, 2JNQ

4. Discussion

Currently, there is a very limited number of inteins suitable for biotechnical applications because such inteins require (i) a most robust splicing activity *in vivo* and *in vitro*, (ii) fast reaction kinetics, (iii) a high tolerance of foreign extein contexts, and (iv) functional reconstruction of their catalytically-active structures from split fragments in order to enable a versatile use of protein splicing [10,14,15]. Extremophiles could be good sources for hunting new robust inteins because of the prevalence of inteins [4,6,7]. Although more than 1500 inteins have been identified, only a dozen of mini-inteins have been biochemically characterized [7]. In the past, biochemical as well as structural studies of uncharacterized inteins have enabled identifying novel robust inteins [15,19]. Here, we determined the three-dimensional structure of a naturally-occurring mini-intein, *Pho*Pol-II intein, and compared it with two other natural mini-inteins in *Pyrococcus horikoshii*. Even though all three mini-inteins share the same HINT fold with an extended β-strand insertion characteristic for inteins from thermophilic organisms [27], the three inteins show distinct differences in helical lengths and loop insertions. The distributed insertion and deletion differences for the entire sequences support the view of specific evolutionary pathways

 doi:10.20944/preprints202105.0161.v1

originating from unique ancestors. Among the three mini-inteins of *Pyrococcus horikoshii*, only *Pho*Pol-II intein requires an elevated temperature for some protein splicing activity *in vitro*, similarly to the *Pab*Pol-II intein [33]. The dependence of active DNA polymerase II production on protein-splicing activity implies that the *Pho*Pol-II intein could play a critical role in regulating the fitness of the organism in response to environmental stimuli, such as temperature changes because DNA polymerase II is an essential enzyme. With this strategy, mini-inteins like *Pho*Pol-II intein could avoid their removal under certain conditions [5].

Mini-inteins lack HEN domains, either because they have been lost during evolution and/or because they have not yet been invaded by a homing endonuclease [13]. The closely related archaea *Pyrococcus horikoshii* and *Pyrococcus abyssi* contain 14 inteins in their genomes. Interestingly, a RadA intein with high splicing activity at room temperature is absent in *Pyrococcus abyssi*, whereas both CDC21 and Pol-II mini-inteins are present in both *P. horikoshii* and *P. abyssi*. The *Pho*RadA intein, which is highly capable of splicing at ambient temperatures, might face a higher elimination pressure in hyperthermophilic organisms than the other mini-inteins, which require elevated temperatures for efficient splicing activity.

Most mini-inteins from extreme halophiles are not halo-tolerant but halo-obligatory inteins, meaning they require high salinity for protein-splicing activity. Mini-inteins must acquire a certain mutualism during evolution to become persistent by providing certain post-translational benefits to the host to ensure survival under specific cellular and environmental conditions [5]. Indeed, salt-dependent inteins seem to give some advantages to the host organism [37]. Because of this, inteins that are highly conserved across a wide phylogenetic distribution might not be the most promising candidates for identifying novel robust mini-inteins. Such highly conserved mini-inteins residing at conserved insertion sites likely to have developed a significant degree of mutualism in physiological conditions, eventually avoiding their removal under certain environmental or cellular conditions. Demonstrated examples include an elevated temperature for hyperthermophiles and a high salinity for extreme halophiles. Uncharacterized mini-inteins might provide some benefits to the host with other unknown regulatory functions as found in RadA intein [38]. Naturally occurring mini-inteins that are poorly conserved among different species and exist sporadically in the genome might be better candidates for further biochemical and structural characterizations, although they might still have unknown reasons/functions to be persistent in their host organisms in physiological condition, such as the mutualisms between mini-inteins and host proteins [39].

An alternative way to identify mini-inteins with efficient splicing activity could be to artificially engineer inteins without HEN domains by removing them. There are more inteins that contain HEN domains than there are naturally-occurring mini-inteins, judging from intein sizes [7,39]. Whereas some inteins are fully capable of catalyzing protein splicing without the HEN domain, others have developed a mutualism with the HEN domain, requiring it for efficient protein-splicing activity [9,14,15,16]. Unfortunately, it is unknown what makes their protein-splicing activities dependent on the presence of a HEN domain. Thus, this approach still requires tedious experimental trials, including constructing several deletion variants [9,14,15,16].

Since only a few three-dimensional structures containing HEN domains exist, further structural elucidation of inteins, particularly those with HEN domains, will provide a better understanding of the structural basis of mutualistic relationship between the HINT and HEN domains. Such a better understanding could lead to the ability to expand the repertoire of promising mini-inteins for developing new biotechnological tools in a rational and predictive way.

Supplementary Materials:

Author Contributions: Conceptualization, HI; methodology, HI, HMB, and MKH; validation, MKH, HMB, and HI; formal analysis, MKH, HMB and HI; investigation, MKH, HMB, and HI;

writing—original draft preparation, MKH and HI; writing—review and editing, MKH, HMB and HI; visualization, MKH and HI; supervision, HI; project administration, HI; funding acquisition, HI. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded in part by the Academy of Finland, decision numbers:137995, 277335 and Novo Nordisk Foundation (NNF17OC0025402 to H. M. B., NNF17OC0027550 to H. I.) and Biocenter Finland, HiLIFE-INFRA and FINStruct for crystallization and NMR facilities at the Institute of Biotechnology.

Data Availability Statement: The coordinate and structure factor of *Pho*Pol-II intein have been deposited to the Protein Data Bank with accession number 7OEC.

Acknowledgments: We thank C. Albert, T. Ayupov L. Krumwiede, and Dr. V. Manole for technical help in protein and plasmid preparations and assay. We thank Prof. A. Wlodawer for his help in structural analysis. We thank Dr. V. Manole for technical help at the crystallization facility.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Paulus, H. Protein splicing and related forms of protein autoprocessing. *Annu. Rev. Biochem*. **2000**, *69*, 447-496.
2. Perler, FB; Davis, EO; Dean, GE; Gimble, FS; Jack, WE; Neff, N;Noren, CJ; Thorner, J; Belfort, M. Protein splicing elements: inteins and exteins-a definition of terms and recommended nomenclature. *Nucleic Acids Res*. **1994**, *22*, 1125-1127. doi:10.1093/nar/22.7.1125
3. Gogarten, JP; Senejani, AG; Zhaxybayeva, O; Olendzenski, L; Hilario, E. Inteins: structure, function, and evolution. *Annu Rev Microbiol*. **2002**, *56*, 263-287. doi:10.1146/annurev.micro.56.012302.160741
4. Novikova, O; Topilina, N; Belfort, M. Enigmatic distribution, evolution, and function of inteins. *J Biol Chem*. **2014**, *289*,14490-14497. doi:10.1074/jbc.R114.548255
5. Belfort, M. Mobile self-splicing introns and inteins as environmental sensors. *Curr Opin Microbiol*. **2017**, *38*, 51-58. doi:10.1016/j.mib.2017.04.003
6. Pietrokovski, S. Intein spread and extinction in evolution. *Trends Genet*. **2001**, *17*, 465-472. doi:10.1016/s0168-9525(01)02365-4
7. Novikova, O; Jayachandran, P; Kelley, DS; Morton, Z; Merwin, S; Topilina, NI; Belfort, M. Intein Clustering Suggests Functional Importance in Different Domains of Life. *Mol Biol Evol*. **2016**, *33*, 783-799. doi:10.1093/molbev/msv271
8. Aranko, AS; Oeemig, JS; Kajander, T; Iwaï, H. Intermolecular domain swapping induces intein-mediated protein alternative splicing. *Nat Chem Biol*. **2013**, *9*, 616-622. doi:10.1038/nchembio.1320
9. Iwaï, H; Mikula, KM; Oeemig, JS; Zhou, D; Li, M; Wlodawer, A. Structural Basis for the Persistence of Homing Endonucleases in Transcription Factor IIB Inteins. *J Mol Biol*. **2017**, *429*, 3942-3956. doi:10.1016/j.jmb.2017.10.016
10. Topilina, NI; Mills, KV. Recent advances in in vivo applications of intein-mediated protein splicing. *Mob DNA*. **2014**, *5*, 5. doi: 10.1186/1759-8753-5-5.
11. Nanda, A; Nasker, SS; Mehra, A; Panda, S; Nayak, S. Inteins in Science: Evolution to Application. *Microorganisms*. **2020**, *8*, 2004. doi:10.3390/microorganisms8122004
12. Ciragan, A; Aranko, AS; Tascon, I; Iwaï, H. Salt-inducible Protein Splicing in cis and trans by Inteins from Extremely Halophilic Archaea as a Novel Protein-Engineering Tool. *J Mol Biol*. **2016**, *428*, 4573-4588. doi: 10.1016/j.jmb.2016.10.006.
13. Gogarten, JP; Hilario, E. Inteins, introns, and homing endonucleases: recent revelations about the life cycle of parasitic genetic elements. *BMC Evol Biol*. **2006**, *6*, 94. doi:10.1186/1471-2148-6-94
14. Aranko, AS; Oeemig, JS; Zhou, D; Kajander, T; Wlodawer, A; Iwaï, H. Structure-based engineering and comparison of novel split inteins for protein ligation. *Mol. Biosyst*. **2014**, *10*, 1023–1034. doi: 10.1039/C4MB00021H
15. Pinto, F; Thornton, EL; Wang, B. An expanded library of orthogonal split inteins enables modular multi-peptide assemblies. *Nat Commun*. **2020**, *11*, 1529. doi:10.1038/s41467-020-15272-2
16. Hiraga, K; Derbyshire, V; Dansereau, JT; Van Roey, P; Belfort, M. Minimization and stabilization of the Mycobacterium tuberculosis recA intein. *J Mol Biol*. **2005**, *354*, 916-926. doi: 10.1016/j.jmb.2005.09.088
17. Oeemig, JS; Zhou, D; Kajander, T; Wlodawer, A; Iwaï, H. NMR and crystal structures of the Pyrococcus horikoshii RadA intein guide a strategy for engineering a highly efficient and promiscuous intein. *J. Mol. Biol.* **2012**, *412*, 85-99
18. Beyer, HM; Mikula, KM; Kudling, TV; Iwaï, H. Crystal structures of CDC21-1 inteins from hyperthermophilic archaea reveal the selection mechanism for the highly conserved homing endonuclease insertion site. *Extremophiles*. **2019**, *23*, 669-679. doi:10.1007/s00792-019-01117-4
19. Ellilä, S; Jurvansuu, JM; Iwaï, H. Evaluation and comparison of protein splicing by exogenous inteins with foreign exteins in Escherichia coli. *FEBS let.* **2011**, *585*, 3471–3477. doi: 10.1016/j.febslet.2011.10.005
20. Guerrero, F; Ciragan, A; Iwaï, H. Tandem SUMO fusion vectors for improving soluble protein expression and purification. *Protein Expr Purif*. **2015**, *116*, 42-49. doi:10.1016/j.pep.2015.08.019
21. Kabsch, W. XDS. *Acta Crystallogr D Biol Crystallogr*. **2010**, *66*(Pt 2), 125-132. doi:10.1107/S0907444909047337

22. Waterhouse, A; Bertoni, M; Bienert, S; Studer, G; Tauriello, G; Gumienny, R; Heer, FT; de Beer, TAP; Rempfer, C; Bordoli, L; Lepore, R; Schwede, T. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **2018**, *46*(W1), W296-W303. doi: 10.1093/nar/gky427.
23. Langer, G; Cohen, SX; Lamzin, VS; Perrakis, A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat Protoc*. **2008**, *3*, 1171-1179. doi:10.1038/nprot.2008.91
24. Emsley, P; Lohkamp, B; Scott, WG; Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr*. **2010**, *66*(Pt 4), 486-501. doi:10.1107/S0907444910007493
25. Adams, PD; Grosse-Kunstleve, RW; Hung, LW; Ioerger, TR; McCoy, AJ; Moriarty, NW; Read, RJ; Sacchettini, JC; Sauter, NK; Terwilliger, TC. PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr D Biol Crystallogr*. **2002**, *58*, 1948-1954. doi:10.1107/s0907444902016657
26. Du, Z; Liu, J; Albracht, CD; Hsu, A; Chen, W; Marieni, MD; Colelli, KM; Williams, JE; Reitter, JN; Mills, KV; Wang, C. Structural and mutational studies of a hyperthermophilic intein from DNA polymerase II of Pyrococcus abyssi. *J Biol Chem.* **2011**, *286*, 38638-38648. doi: 10.1074/jbc.M111.290569
27. Aranko, AS; Wlodawer, A; Iwaï, H. Nature's recipe for splitting inteins. *Protein Eng. Des. Sel.* **2014**, *27*, 263–271.
28. Razvi, A; Scholtz, JM. Lessons in stability from thermophilic proteins. *Protein Sci.* **2006**, 15, 1569-78. doi: 10.1110/ps.062130306..
29. Brünger, AT. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*. **1992**, *355*(6359), 472-475. doi:10.1038/355472a0
30. Karplus, PA; Diederichs, K. Linking crystallographic model and data quality. *Science*. **2012**, *336*, 1030-1033. doi:10.1126/science.1218231
31. Johnson, MA; Southworth, MW; Herrmann, T; Brace, L; Perler, FB; Wüthrich, K. NMR structure of a KlbA intein precursor from Methanococcus jannaschii. *Protein Sci.* **2007**, *16*, 1316-28. doi: 10.1110/ps.072816707
32. González, JM; Masuchi, Y; Robb, FT; Ammerman, JW; Maeder, DL; Yanagibayashi, M; Tamaoka, J; Kato, C. Pyrococcus horikoshii sp. nov., a hyperthermophilic archaeon isolated from a hydrothermal vent at the Okinawa Trough. *Extremophiles*. **1998**, *2*, 123-130. doi: 10.1007/s007920050051.
33. Mills, KV; Manning, JS; Garcia, AM; Wuerdeman, LA. Protein splicing of a Pyrococcus abyssi intein with a C-terminal glutamine. *J Biol Chem*. **2004**, *279*, 20685-91. doi: 10.1074/jbc.M400887200
34. Erauso, G; Reysenbach, AL; Godfroy, A; Meunier, JR; Crunp, B; Partensky, F; Baross, JA; Marteinsson, V; Barbier, G; Pace, NR; Prieur, D. Pyrococcus abyssi sp. nov., a new hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent. *Arch. Microbiol*. **1993**, *160*, 338–349. doi: 10.1007/BF00252219
35. Huber, H; Stetter, KO. Thermoplasmatales. In: Dworkin, M; Falkow, S; Rosenberg, E; Schleifer, KH; Stackebrandt, E. (eds) The Prokaryotes. Springer, New York, NY, USA, **2006**; pp. 101-112. doi: 10.1007/0-387-30743-5_7
36. Tsoka, S; Simon, D; Ouzounis, CA. Automated metabolic reconstruction for Methanococcus jannaschii. *Archaea*. **2004**, *1*, 223-229. doi: 10.1155/2004/324925
37. Naor, A; Altman-Price, N; Soucy, SM; Green, AG; Mitiagin, Y; Turgeman-Grott, I; Davidovich, N; Gogarten, JP; Gophna, U. Impact of a homing intein on recombination frequency and organismal fitness. *Proc Natl Acad Sci U S A*. **2016**, *113*, E4654-E4661. doi:10.1073/pnas.1606416113
38. Topilina, NI; Novikova, O; Stanger, M; Banavali, NK; Belfort, M. Post-translational environmental switch of RadA activity by extein-intein interactions in protein splicing. *Nucleic Acids Res*. **2015**, *43*, 6631-6648. doi:10.1093/nar/gkv612
39. Green, CM; Novikova, O; Belfort, M. The dynamic intein landscape of eukaryotes. *Mob DNA*. **2018**, *9*:4. doi:10.1186/s13100-018-0111-x