biomolecules

MDPI

Article

The Repeating, Modular Architecture of the HtrA Proteases

Matthew Merski [1,*], Sandra Macedo-Ribeiro [2], Rafal M. Wieczorek [3] and Maria W. Górna [1,*]

[1] Structural Biology Group, Biological and Chemical Research Centre, Faculty of Chemistry, University of Warsaw, Żwirki i Wigury 101, 02-089 Warsaw, Poland
[2] Instituto de Investigação e Inovação em Saúde and Instituto de Biologia Molecular e Celular (IBMC), Universidade do Porto, 4200-135 Porto, Portugal; sribeiro@ibmc.up.pt
[3] Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland; wieczorek@chem.uw.edu.pl
* Correspondence: merski@gmail.com (M.M.); mgorna@chem.uw.edu.pl (M.W.G.); Tel.: +48-225-526-642 (M.M.)

Abstract: A conserved, 26-residue sequence [AA(X_2)[A/G][G/L](X_2)GDV[I/L](X_2)[V/L]NGE(X_1)V(X_6)] and corresponding structure repeating module were identified within the HtrA protease family using a non-redundant set (N = 20) of publicly available structures. While the repeats themselves were far from sequence perfect, they had notable conservation to a statistically significant level. Three or more repetitions were identified within each protein despite being statistically expected to randomly occur only once per 1031 residues. This sequence repeat was associated with a six stranded antiparallel β-barrel module, two of which are present in the core of the structures of the PA clan of serine proteases, while a modified version of this module could be identified in the PDZ-like domains. Automated structural alignment methods had difficulties in superimposing these β-barrels, but the use of a target human HtrA2 structure showed that these modules had an average RMSD across the set of structures of less than 2 Å (mean and median). Our findings support Dayhoff's hypothesis that complex proteins arose through duplication of simpler peptide motifs and domains.

Keywords: HtrA protease; protein repeat; PA clan; serine protease; protein evolution

Citation: Merski, M.; Macedo-Ribeiro, S.; Wieczorek, R.M.; Górna, M.W. The Repeating, Modular Architecture of the HtrA Proteases. *Biomolecules* **2022**, *12*, 793. https://doi.org/10.3390/biom12060793

Academic Editor: Jian Zhang

Received: 24 May 2022
Accepted: 4 June 2022
Published: 7 June 2022

1. Introduction

Generally, the arrangement of amino acids in proteins is seemingly random (complex), although exceptions exist where notable patterns can be discerned in the amino acid sequence, such as low-complexity proteins [1] or protein repeats [2–4]. Proteins also usually adopt distinct three-dimensional structures and a wide variety of these have been reported in the public repository of the Protein Data Bank (PDB) [5]. A combination of elements (sequence, secondary structure, fold, and three-dimensional structure) comprise the architecture of a protein. However, the three-dimensional structure itself tends to be the most conserved aspect of the protein as both the sequence and function of the protein evolve much more quickly [6], although it has been suggested that the folds themselves are fossils [7] of the early, Archean proteins which may have evolved before the appearance of the last universal common ancestor (LUCA) [8]. Five decades ago, based on the earliest protein structures, it was hypothesized that these primitive peptides formed oligomeric groups in solution which eventually fused into single peptide transcripts to give rise to the early, modern proteins which then over time gradually diverged into more complex forms [9–13]. This process of oligomerization followed by fusion has also been suggested to have given rise to repeat proteins, which are composed of a set of repeating structures and sequences of 20–60 amino acids in length, which may have, over time, evolved into complex, globular proteins (Figure 1) [2,14,15]. While there are a number of well-known repeat protein types, they have generally received less researcher attention than globular proteins that have more complex structural architectures despite estimates suggesting that about a quarter of all known proteins have at least some repeat protein character [16]. This raises the obvious question as to what is obscuring the presence of all these expected

protein repeats in structural databases such as the PDB, especially given the possibility that the early, ancestral proteins were all at least repeat-like [17–19].

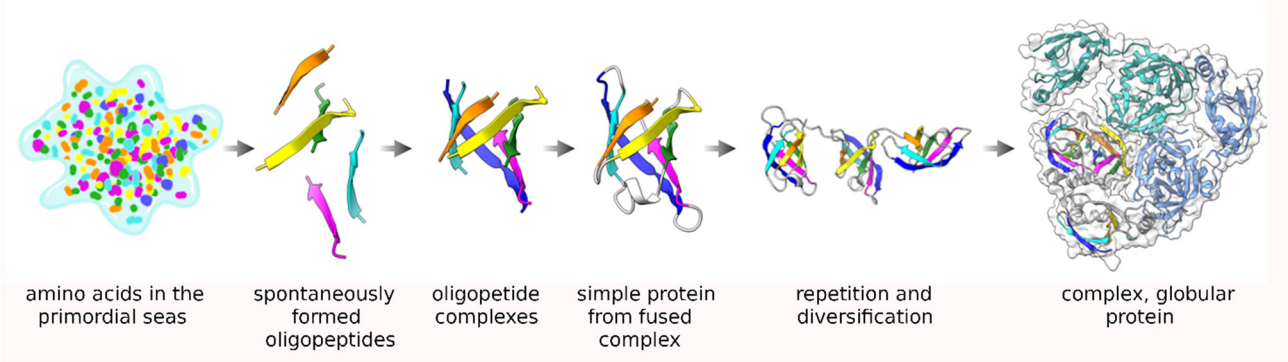

Figure 1. An illustration of Dayhoff's hypothesis about the origin of proteins [9,17]. From left to right, starting from individual, spontaneously formed amino acids in the Archaean seas, short oligopeptides formed spontaneously which then organized into homogenous complexes and eventually fused into a single transcript module, probably after being encoded in the genome. Duplication and repetition of these modules along with drift in their sequence and function eventually gave rise to complex, globular proteins.

Protein families that are widely distributed across the three kingdoms of life are likely to have roots deep in evolutionary time [8,15], possibly even as far back as the Archean, pre-LUCA period, and may be, in essence, representatives of such preserved fossil architectures. One such protein family could be the HtrA family of proteases. The high temperature requirement A (HtrA) proteases are stress response, housekeeping proteases widely distributed throughout nature [20]. Notable examples of this family include DegP [21] and DegS [22] in prokaryotes, Deg1 in plants [23], and HtrA2 in humans [24–27]. Structurally, HtrA proteases are members of the PA clan of serine proteases (including such notable examples as chymotrypsin A and thrombin) which contain a pair of six-stranded β-barrels [28]. HtrA proteases additionally have one or more C-terminal PDZ-like domains [25,29], an 80–100 amino acid long protein interaction domain found in many different protein families [30]. In prokaryotes, DegP [21] forms large 12- and 24-mer complexes while DegS [22] exists as a simple trimer. In humans, the chromosomally encoded HtrA2 protease, linked to Parkinson's disease [24], functions as a housekeeping protease within the mitochondria [26]. Damage to the mitochondrial membranes results in leakage of HtrA2 into the cytoplasm, where it digests peptide inhibitors of apoptosis leading to cell death [31]. HtrA2 has been shown to have an unusually high melting temperature [26] and to preferably cleave unfolded substrate ensembles [32]. HtrA2 is maintained in a resting closed state and its activation mechanism is a set of sequential steps that are initiated by the binding of a hydrophobic motif to the PDZ-like domain, followed by exposure of the substrate binding site on the protease domain and activation of the proteolytic activity.

We have recently reported [33] a survey of all the known protein sequences using a self-homology detection method based on DOTTER [34]. This allowed us to identify a number of protein families which had a notable amount of self-similarity, including the HtrA protease family. More detailed examination confirmed the initial detection of the repeating amino acid sequence and we were able to correlate the sequence repeats with a six-strand antiparallel β-barrel structure that occurred at least three times in the monomeric structure of the protease (twice in the protease domain, a feature of the PA clan of serine proteases [28]) (Figure 2) and once in each PDZ-like domain. These results suggest that the PDZ-like domain evolved from repetition of this basic barrel structure in the PA clan serine proteases.

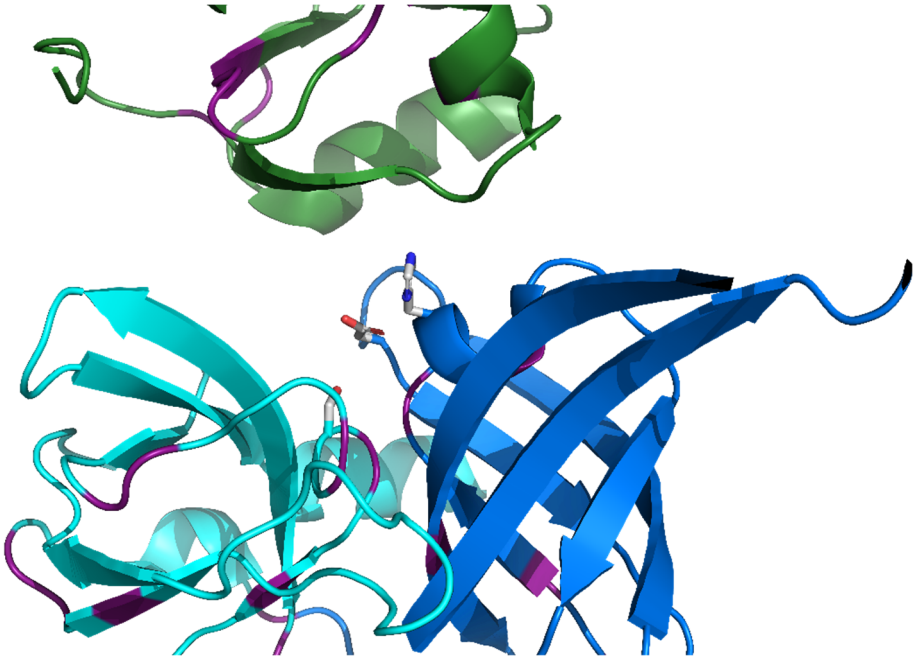

Figure 2. The active site in the HtrA proteases is separate between the modules. Cartoon diagram of human HtrA2 (PDB ID 5m3n [26]) showing the N-terminal protease (blue), C-terminal protease (cyan) and PDZ-like (green) modules. The catalytic triad of His198, Asp 228, and Ser306 are shown as sticks with light grey carbon, blue nitrogen, and red oxygen atoms. Those residues which correspond to conserved canonical repeat residues are indicated in purple (Figure S1).

2. Materials and Methods

As previously reported, all known proteins (UniRef90 [35]) were examined for self-homology using a modified version of DOTTER [33,34]. This analysis found a number of proteins with significant self-homology that were in a protease Do-like cluster. HtrA proteases were then collected from the PDB [5]. Short sequences (less than 200 residues) were removed and the remainder were filtered at the 90% sequence identity threshold with CD-HIT [36], leaving 20 unique structures (i.e., 2zle, 2z9i, 3gdv, 3nzi, 3pv5, 3qo6, 4a9g, 4fln, 4ic5, 4ic6, 4ri0, 4ynn, 5fht, 5ilb, 5jyk, 5t69, 5zvj, 6jjo, 6z05, 7co3). The locations of probable sequence repeats were identified by reverse calculation of the DOTTER plots of these protein sequences where each residue was assigned its maximal self-homology score. The high scoring regions from all the proteins were separated and the frequency of each amino acid at each position was calculated; those positions which had strong biases towards a single, or a pair of similar, amino acids were noted. This process identified a 26-residue repeating sequence. Multiple sequence alignment of these repeats with MUSCLE [37] (Figure S1) helped to clarify the repeated sequence in which 13 of the 26 positions were conserved, specifically [AA(X_2)[A/G][G/L](X_2)GDV[I/L](X_2)[V/L]NGE(X_1)V(X_6)], which can also be represented as AA–[A/G][G/L]–GDV[I/L]–[V/L]NGE-V——. This repeat was then used for further sequence searches.

The statistical significance of this sequence repeat was verified by comparison of the repeat pattern to a randomly generated sequence. A score estimate for the random sequence was defined by a binomial (Bernoulli) model. The random chance for success at each position was defined as the probability for the expected amino acid at that position for each of the 13 defined positions. More explicitly, suppose a set of coin flips (Bernoulli trials) such that each event will have 13 trials and each of those trials have a probability of success defined by the natural frequency of the amino acids that are acceptable in that position. The score for each event is defined by the number of successes that occur in the event. For each success, a score of 1/P(x) where P(x) is the probability of finding an acceptable amino acid in that position is given while failures get a score of 1. For positions with 2 acceptable

amino acids, the probability is the sum of the two amino acid frequencies. This allowed a success score to be defined for any actual sequence as

$$\text{Score} = \Pi(T)$$

where the score value for success T(S) in any position is

$$T(S) = 1/P(x)$$

and for failure T(F) is

$$T(F) = 1$$

Failures are given a score of one in order to not modify the score (i.e., a sequence with no matches to the defined repeat sequence received a score of 1). To further verify this model, a 9996060-residue length of random amino acid sequence was generated by the Sequence Manipulation Suite [38] and each 26-residue sequence unit was compared to the repeat sequence to generate an estimate of the probability for the repeat sequence to appear randomly (Figures S2 and S3). Sequences with a score greater than 90,000 appeared with a frequency of less than 1 per 1000 residues, in agreement with the theoretical binomial model.

The repeat sequence was then identified in a multiple sequence alignment (MUSCLE [37]) in the set of protein structures (Figure S1). These were comprised of a pair of six-stranded barrels in the protease domain and a partial, four-stranded barrel in each PDZ-like domain (Figure S4). The PDB structures were divided into the individual barrel structures and compared by structural alignment in PyMol [39]. The sequence repeats did not share a common secondary structure. However, a shared common antiparallel β-barrel structure could be identified within the structures, which was associated with the sequence repeats (Figure S5). While generally, the alignment of the domains to each other was poor using PyMol, the three domains from a specific human HtrA2 protease (PDB ID 5m3n) could all be superimposed when aligned in PyMol. The N terminal protease, C terminal protease, and PDZ domain modules from all the other example HtrA proteases could then be aligned to the equivalent module from the 5m3n structure and a good superimposition was achieved (Figure S6). The modules were also compared by TM-align [40].

3. Results

Using a modified form of DOTTER to analyze protein self-homology, the presence of notable self-homology was detected in the HtrA proteases. Reverse calculation of the homology plots produced by DOTTER [34] using HtrA protease sequences from the PDB allowed the detection of a putative 26 amino acid repeat [AA(X_2)[A/G][G/L](X_2)GDV[I/L](X_2)[V/L]NGE(X_1)V(X_6)] and alignment of the repeating regions with MUSCLE [37] produced an initial estimate of the repeating sequence with 13 of the 26 positions being definable (Figures 3 and S1). Comparison to a random sequence and theoretical comparison to a Bernoulli model for cumulative probability (Figures S2 and S3) suggested that high scoring matches (typically five or more matching positions) should be relatively uncommon, occurring randomly only once every 1031 residues, which is much less than once per protein as the average HtrA protease monomer is 350–450 residues long [41]. Examination of the sequence unique HtrA proteases (90% ID) clearly identified three or more of these repeats in each monomer, corresponding roughly to two in the protease domain and one in each PDZ-like domain (Figure S4) (some HtrA proteases such as DegP have two copies of the PDZ-like domain [30]). The protease and PDZ-like domains within the HtrA proteases did not have significant self-homology (mean = 18.2%, median = 15.3% for protease and PDZ-like 1 domains, N = 19 proteins (Table S1)). Only one protein, *Legionella pneumophila* DegQ (PDB ID 4ynn) [42], had greater than 30% identity between its protease and PDZ-like domains (34.0% ID). Low shared sequence identity has been previously noted in PDZ domains [43]. Further visual inspection of the structures identified a common alternating

antiparallel six-strand β-barrel module in the structures (Figures 4 and S5). There is a shift in register in the PDZ-like module in which the first beta strand occurs outside of the barrel structure and the barrel remains unclosed as it contains only four additional strands (Figure S5). The sixth strand is also rotated out of the structure or deleted, depending on the species. The alternating anti-parallel pattern present in the protease modules is maintained in the PDZ-like module but formally reversed as the designation of positive and negative strands is arbitrary. RMSD comparison of these isolated structures using PyMol found a poor structural similarity among the β-barrels with a mean RMSD of 4.1 Å between the two protease domains and mean RMSD values of 8.8 Å and 9.2 Å between the N-terminal protease module or the C-terminal protease module and the first PDZ-like domain, respectively (Table S2). The superposition of the PDZ-like and protease modules was poor but their small size makes the RMSD values appear better than they are. Analysis by TM-align [40] suggested good agreement between the protease domains (TM scores = 0.552 (mean), 0.549 (median) for the protease modules and 0.652 (mean), 0.638 (median) for the PDZ-like modules; RMSD = 2.80 Å (mean), 2.78 Å (median) for the protease domains and 2.28 Å (mean), 2.22 Å (median) for the PDZ-like domains). However, there was poor correspondence when the protease modules were compared to the PDZ-like modules or vice-versa (TM-score = 0.298 (mean), 0.299 (median); RMSD = 3.72 Å (mean), 3.70 Å (median)) (Table S3). The modules from one human HtrA2 structure (PDB ID 5m3n [26]) could be structurally aligned after manual examination (Figure 4, Table S3). The aligned modules from this specific PDB structure could then be used as "targets" for the corresponding modules in the other proteins. When this was done, the structural differences were generally minimized and good structural alignments could be achieved (mean = 2.9 Å, 2.9 Å, 1.9 Å; median = 2.5 Å, 2.8 Å, 1.7 Å for the N-terminal protease, C-terminal protease, and PDZ-like domain modules, respectively) (Table S3).

```
                          ag     l  l
Repeat sequence       AA--gl--GDVi--vNGE-v------

4fln (A. thaliana Deg2)
N-protease     142    STGSAFMIGDGKLLTNAHCVEHDTQV   167
C-protease-A   261    AAINPGNSGGPAFNDQGECIGVAFQV   286
PDZ-like 1-A   330    LENPALRECLKVPTNEGVLVRRVEPT   355
PDZ-like 1-B   358    ASKV-LKEGDVIVSFDDLHVGCEGTV   382

4ri0 (H. sapiens HtrA3)
C-protease-A   237    KKLPVLLLGHSADLRPGEFVVAIGSP   262
C-protease-B   298    AIINYGNAGGPLVNLDGEVIGINTLK   323
PDZ-like 1     397    SQRGGIQDGDIIVKVNGRPLVDSSEL   422

5fht (H. sapiens HtrA2)
C-protease-A   105    EPLPTLPLGRSADVRQGEFVVAMGSP   130
C-protease-B   166    AAIDFGNAGGPLVNLDGEVIGVNTMK   191
PDZ-like 1     269    AHRAGLRPGDVILAIGEQMVQNAEDV   294

4ynn (L. pneumophila DegQ)
N-protease     120    KNLKSIVIGDSDKLEVGDYVVAIGNP   145
C-protease-A   156    SATFGIVSALKRSDLNIEGVENFIQT   181
C-protease-B   183    AAINPGNAGGALVNAKGELIGINTAI   208
PDZ-like 1     280    AQLAGLKSGDVIVQINDTKITQATQV   305
PDZ-like 2     384    GWRAGLRPGDIIISANKTPVKDIKSL   407

Repeat position       12345678901234567890123456
                                1         2
```

Figure 3. Identification of the sequence repeats in the HtrA proteases. The canonical sequence [AA(X_2)[A/G][G/L](X_2)GDV[I/L](X_2)[V/L]NGE(X_1)V(X_6)] is shown on top in bold, with an additional residue shown for the four positions which have two possible canonical residues. Residues that match the canonical sequence are highlighted in green. The PDB ID, species, and protein name are given along with the module in which the sequence is located. When a module has two copies of the sequence repeat, the most N-terminal is denoted as A and the other as B. The beginning sequence position (using the PDB numbering) is to the left of the sequence while the ending position is given to the right of the sequence.

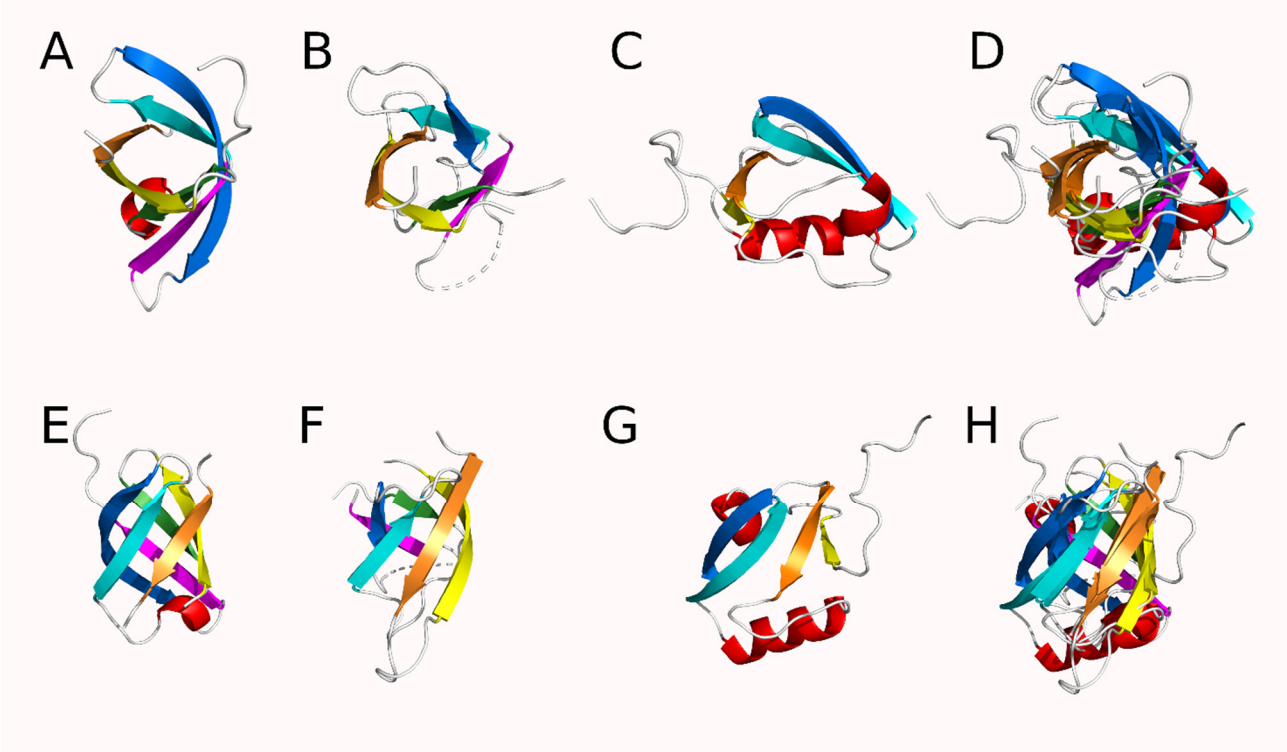

Figure 4. Cartoon diagram of the modules from HhoA, an HtrA protease from *Synechocystis* sp. PCC 6803 (PDB ID 5t69) showing the conserved structures of the HtrA modules (RMSD to PDB ID 7co3: mean = 1.748 Å, median = 1.816 Å). Strands are colored orange, yellow, green, blue, and magenta in order from N to C in the protease modules and in the equivalent spatial position in the PDZ-like module. Helices are colored red and coil regions are white. Top-down views of the (**A**) N-terminal protease module, (**B**) C-terminal protease module, (**C**) PDZ-like module and (**D**) all three modules superimposed. Side views of the (**E**) N-terminal protease module, (**F**) C-terminal protease module, (**G**) PDZ-like module, and (**H**) all three modules superimposed.

4. Discussion

Fundamentally, the HtrA proteases are repeat proteins. A 26-residue sequence repeat associated with an anti-parallel β-barrel structure is clearly identifiable in the HtrA proteases (structures shown in Figures 2 and 4). While the repeating sequences are far from perfect [44], the frequency of matches to the defined canonical sequence is statistically significant (Figures S2 and S3). The individual modules can be structurally aligned to a set of target modules to a good average RMSD (<3 Å) between the β-barrel structural modules within a given protein (Figure 4, Table S3). Repetitions of both sequence and structure in combination with the presence of two copies of a β-barrel in the PA clan of serine proteases [28] (to which HtrA proteases belong) and all three modules having a peptide binding function strongly suggests that the HtrA proteases are the result of a set of repetitions of the ancestral β-barrel module followed by mutation and functional change in the third (by sequence order) module present in the PDZ-like domain(s) [45].

The evolution of the modern HtrA protease structure from an ancestral β-barrel precursor, possibly an Archean, pre-LUCA protease [8], via the PA clan ancestor [28] offers an elegant solution to the problem of the origin of structural complexity of this family of proteases from a simple ancestor as suggested by Dayhoff's hypothesis [9,17]. It is currently unknown if the ancestral module itself was an active protease or if it simply had a peptide binding function and developed into an active protease after the duplication at the origin of the PA clan as the catalytic triad of the HtrA proteases is spread across the two protease modules. For example, in human HtrA2 [25], *E. coli* DegP [46], and *A. thaliana* Deg2 [47], the catalytic serine is found in the C-terminal protease module, while the other two members of

the triad are present in the first module. The PDZ-like module is likely derived from one of these modules as it contains several divergent structural features compared to the protease modules. The protease modules have six strands comprising its β-barrel, while only four of these are present in the PDZ-like domain module [45] along with an additional N-terminal strand which is rotated out of the barrel structure (Figure S5). There are also many PA clan proteases which lack the PDZ-like domain, suggesting that it evolved later. Therefore, while it is not undisputable, it seems likely that the PDZ-like module is a product of the duplication of one of the protease modules rather than the protease being derived from the PDZ-like domain. This may be an incorrect assumption, however, given the amount of lateral gene transfer that occurs in prokaryotes [48,49].

By analyzing the conserved self-homology patterns in the HtrA proteases, we were able to identify the simple, repeating β-barrel architecture present in this family. To the best of our knowledge, this repeating architecture has gone unremarked upon despite the fact that structures of these proteins have been publicly available for 20 years [25,29] and the widely recognized pair of β-barrels present in the PA clan of proteases. This was likely at least partially due to the general difficulty in identifying protein repeats [50–52]. In this specific case, there are one (or two) sequence repeats present in each of the β-barrel structural modules (Figure S1), and a small but notable discrepancy between the sequence and structural repetitions, which would contribute to the difficulty in identifying these repeats. However, a discrepancy between sequence and structural repeats is not uncommon in repeat proteins [53]. Additionally, automated structural alignment methods had difficulty in detecting the similarity between the modules, even after the repeats had been unambiguously identified. The low sequence similarity between the members of the family or the different repeat modules as well as the low sequence conservation within the repeats themselves likely contributed to this detection issue, as did the variability in the assignment of the secondary structures in the protein structure models themselves (Tables S2 and S3). It is also worth noting that several standard structural alignments failed to properly superimpose the protease modules with PDZ-like modules., However, they could be convinced to superimpose the modules when a properly superimposed "target" structure was used (Figure S6, Table S3). We must also note that even the method used here has its limitations. The length is defined by a relatively well-conserved valine at position 26 (Figure 3). However, this valine does not occur as frequently in the data set as the defined canonical residues, and certainly not at the 40% frequency used to define other canonical repeat residues [54], but its presence did help to define the length of the HtrA repeats described here.

Nevertheless, the successful detection of this overlooked repeat architecture in a well-studied family of proteins using self-homology does not imply that this method for repeat detection cannot be further improved. While the method was able to find the sequence repeats fairly easily, they did not correspond to the structural repeat and identification of that required significant human intervention. Even when the β-barrel was recognized, the movement of the first and last strand out of the barrel and the spatial rearrangement of the strands prevented accurate matching of the β-barrels using structural alignment algorithms without human optimization. Clearly, improvements in the automation of the structure search strategies would be beneficial here, since simple removal of the coil regions did not improve detection ability, quite likely due to discrepancies in identification of secondary structural features in the crystallographic models (Table S3). Finally, despite identification of these repeating modules, it still cannot be indisputably determined which module is the most ancestral and which are derived.

Despite these caveats, the conserved, repeating architecture of the HtrA proteases is clearly identifiable in the family. Self-homology analysis was able to identify this architecture which had gone overlooked for decades, a clear success for this method of repeat detection. This repeat architecture shows an elegant method to generate complex protein structures from simple oligopeptide building blocks and might serve to inform protein

engineering efforts. This repeat detection methodology can (and will) be applied to other well-studied protein families and potentially identify their underlying repeat architectures.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/biom12060793/s1, Figure S1: MUSCLE 3.8 Alignment of HtrA proteases, Figure S2: Frequency of HtrA repeats in random protein sequences, Figure S3: Expected protein length of a random sequence matching that score to occur, Figure S4: Identification of the β–barrel structures in the HtrA proteases, Figure S5: Secondary structures of the HtrA modules, Figure S6: Structures of the HtrA modules, Table S1: Domain sequence similarity comparison, Table S2: Unmodified RMSD comparison table, Table S3: β-barrel comparison summary table.

Author Contributions: Conceptualization, M.M.; methodology, M.M.; software, M.M.; formal analysis, M.M. and S.M.-R.; investigation, M.M.; resources, R.M.W. and M.W.G.; data curation, M.M. and M.W.G.; writing—original draft preparation, M.M., S.M.-R., R.M.W. and M.W.G.; writing—review and editing, M.M., S.M.-R., R.M.W. and M.W.G.; visualization, M.M. and M.W.G.; supervision, S.M.-R.; funding acquisition, M.W.G. and R.M.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Science Centre, Poland, grant #2014/15/D/NZ1/00968 to M.W.G and by the European Union's Horizon 2020 research and innovation programme under GA 952334 (PhasAGE) to S.M.-R.

Data Availability Statement: Protein structures analyzed in this work are publicly available in the PDB (www.rcsb.org, accessed 1 April 2022).

Acknowledgments: The authors would like to Dominik Gront for helpful discussions.

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

References

1. Radó-Trilla, N.; Albà, M. Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. *BMC Evol. Biol.* **2012**, *12*, 155. [CrossRef] [PubMed]
2. Andrade, M.A.; Perez-Iratxeta, C.; Ponting, C.P. Protein repeats: Structures, functions, and evolution. *J. Struct. Biol.* **2001**, *134*, 117–131. [CrossRef] [PubMed]
3. Espada, R.; Parra, R.G.; Sippl, M.J.; Mora, T.; Walczak, A.M.; Ferreiro, D.U. Repeat proteins challenge the concept of structural domains. *Biochem. Soc. Trans.* **2015**, *43*, 844–849. [CrossRef]
4. Kajava, A.V. Tandem repeats in proteins: From sequence to structure. *J. Struct. Biol.* **2011**, *179*, 279–288. [CrossRef]
5. Burley, S.K.; Berman, H.M.; Bhikadiya, C.; Bi, C.X.; Chen, L.; Di Costanzo, L.; Christie, C.; Dalenberg, K.; Duarte, J.M.; Dutta, S.; et al. RCSB Protein Data Bank: Biological macromolecular structures enabling research and education in fundamental bi-ology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* **2019**, *47*, D464–D474. [CrossRef]
6. Illergård, K.; Ardell, D.H.; Elofsson, A. Structure is three to ten times more conserved than sequence-A study of structural response in protein cores. *Proteins: Struct. Funct. Bioinform.* **2009**, *77*, 499–508. [CrossRef]
7. Alva, V.; Söding, J.; Lupas, A.N. A vocabulary of ancient peptides at the origin of folded proteins. *eLife* **2015**, *4*, e09410. [CrossRef]
8. Ranea, J.A.G.; Sillero, A.; Thornton, J.M.; Orengo, C.A. Protein superfamily evolution and the last universal common an-cestor (LUCA). *J. Mol. Evol* **2006**, *63*, 513–525. [CrossRef] [PubMed]
9. Eck, R.V.; Dayhoff, M.O. Evolution of the Structure of Ferredoxin Based on Living Relics of Primitive Amino Acid Sequences. *Science* **1966**, *152*, 363–366. [CrossRef]
10. Alva, V.; Lupas, A.N. From ancestral peptides to designed proteins. *Curr. Opin. Struct. Biol.* **2018**, *48*, 103–109. [CrossRef]
11. Broom, A.; Doxey, A.C.; Lobsanov, Y.D.; Berthin, L.G.; Rose, D.R.; Howell, P.L.; McConkey, B.J.; Meiering, E.M. Modu-lar Evolution and the Origins of Symmetry: Reconstruction of a Three-Fold Symmetric Globular Protein. *Structure* **2012**, *20*, 161–171. [CrossRef] [PubMed]
12. Jackson, V.A.; Busby, J.N.; Janssen, B.; Lott, S.; Seiradake, E. Teneurin Structures Are Composed of Ancient Bacterial Protein Domains. *Front. Neurosci.* **2019**, *13*, 183. [CrossRef] [PubMed]
13. Wieczorek, R. On Prebiotic Ecology, Supramolecular Selection and Autopoiesis. *Orig. Life Evol. Biosphere* **2012**, *42*, 445–452. [CrossRef] [PubMed]
14. Söding, J.; Lupas, A.N. More than the sum of their parts: On the evolution of proteins from peptides. *BioEssays* **2003**, *25*, 837–846. [CrossRef]
15. Lupas, A.N.; Ponting, C.; Russell, R.B. On the Evolution of Protein Folds: Are Similar Motifs in Different Protein Folds the Result of Convergence, Insertion, or Relics of an Ancient Peptide World? *J. Struct. Biol.* **2001**, *134*, 191–203. [CrossRef]
16. Marcotte, E.; Pellegrini, M.; Yeates, T.; Eisenberg, D. A census of protein repeats. *J. Mol. Biol.* **1999**, *293*, 151–160. [CrossRef] [PubMed]

17. Romero, M.L.R.; Rabin, A.; Tawfik, D.S. Functional Proteins from Short Peptides: Dayhoff's Hypothesis Turns 50. *Angew. Chem. Int. Ed.* **2016**, *55*, 15966–15971. [CrossRef]
18. Laurino, P.; Tóth-Petróczy, A.; Meana-Pañeda, R.; Lin, W.; Truhlar, D.; Tawfik, D.S. An Ancient Fingerprint Indicates the Common Ancestry of Rossmann-Fold Enzymes Utilizing Different Ribose-Based Cofactors. *PLOS Biol.* **2016**, *14*, e1002396. [CrossRef] [PubMed]
19. Galpern, E.A.; Freiberger, M.I.; Ferreiro, D.U. Large Ankyrin repeat proteins are formed with similar and energetically fa-vorable units. *PLoS ONE* **2020**, *15*, e0233865. [CrossRef]
20. Clausen, T.; Kaiser, M.; Huber, R.; Ehrmann, M. HTRA proteases: Regulated proteolysis in protein quality control. *Nat. Rev. Mol. Cell Biol.* **2011**, *12*, 152–162. [CrossRef]
21. Ortega, J.; Iwanczyk, J.; Jomaa, A. *Escherichia coli* DegP: A Structure-Driven Functional Model. *J. Bacteriol.* **2009**, *191*, 4705–4713. [CrossRef]
22. Sohn, J.; Grant, R.A.; Sauer, R.T. Allostery Is an Intrinsic Property of the Protease Domain of DegS: Implications for Enzyme Function and Evolution. *J. Biol. Chem.* **2010**, *285*, 34039–34047. [CrossRef]
23. Kley, J.; Schmidt, B.; Boyanov, B.; Stolt-Bergner, P.C.; Kirk, R.; Ehrmann, M.; Knopf, R.R.; Naveh, L.; Adam, Z.; Clausen, T. Structural adaptation of the plant protease Deg1 to repair photosystem II during light exposure. *Nat. Struct. Mol. Biol.* **2011**, *18*, 728–731. [CrossRef] [PubMed]
24. Hegde, R.; Srinivasula, S.M.; Zhang, Z.J.; Wassell, R.; Mukattash, R.; Cilenti, L.; DuBois, G.; Lazebnik, Y.; Zervos, A.S.; Fernandes-Alnemri, T.; et al. Identification of Omi/HtrA-2 as a mitochondrial apoptotic serine protease that dis-rupts inhibitor of apoptosis protein-caspase interaction. *J. Biol. Chem.* **2002**, *277*, 432–438. [CrossRef]
25. Li, W.Y.; Srinivasula, S.M.; Chai, J.J.; Li, P.W.; Wu, J.W.; Zhang, Z.J.; Alnemri, E.S.; Shi, Y.G. Structural insights into the pro-apoptotic function of mitochondrial serine protease HtrA2/Omi. *Nat. Struct. Biol.* **2002**, *9*, 436–441. [CrossRef]
26. Merski, M.; Moreira, C.; Abreu, R.M.; Ramos, M.J.; Fernandes, P.; Martins, L.M.; Pereira, P.; Macedo-Ribeiro, S. Molecular motion regulates the activity of the Mitochondrial Serine Protease HtrA2. *Cell Death Dis.* **2017**, *8*, e3119. [CrossRef]
27. Zurawa-Janicka, D.; Jarzab, M.; Polit, A.; Skorko-Glonek, J.; Lesner, A.; Gitlin, A.; Gieldon, A.; Ciarkowski, J.; Glaza, P.; Lubomska, A.; et al. Temperature-induced changes of HtrA2(Omi) protease activity and structure. *Cell Stress Chaperones* **2012**, *18*, 35–51. [CrossRef] [PubMed]
28. Di Cera, E. Serine Proteases. *Iubmb Life* **2009**, *61*, 510–515. [CrossRef]
29. Krojer, T.; Garrido-Franco, M.; Huber, R.; Ehrmann, M.; Clausen, T. Crystal structure of DegP (HtrA) reveals a new prote-ase-chaperone machine. *Nature* **2002**, *416*, 455–459. [CrossRef]
30. Lee, H.-J.; Zheng, J.J. PDZ domains and their binding partners: Structure, specificity, and modification. *Cell Commun. Signal.* **2010**, *8*, 8. [CrossRef] [PubMed]
31. Martins, L.M.; Turk, B.E.; Cowling, V.; Borg, A.; Jarrell, E.T.; Cantley, L.C.; Downward, J. Binding specificity and regula-tion of the serine protease and PDZ domains of HtrA2/Omi. *J. Biol. Chem.* **2003**, *278*, 49417–49427. [CrossRef]
32. Toyama, Y.; Harkness, R.W.; Kay, L.E. Structural basis of protein substrate processing by human mitochondrial high-temperature requirement A2 protease. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2203172119. [CrossRef] [PubMed]
33. Merski, M.; Młynarczyk, K.; Ludwiczak, J.; Skrzeczkowski, J.; Dunin-Horkawicz, S.; Górna, M.W. Self-analysis of repeat proteins reveals evolutionarily conserved patterns. *BMC Bioinform.* **2020**, *21*, 1–17. [CrossRef] [PubMed]
34. Sonnhammer, E.L.L.; Durbin, R. A dot-matrix program with dynamic threshold control suited for genomic DNA and pro-tein sequence analysis (Reprinted from Gene Combis, vol 167, pg GC1-GC10, 1996). *Gene* **1995**, *167*, Gc1–Gc10. [CrossRef]
35. Bateman, A.; Martin, M.J.; Orchard, S.; Magrane, M.; Alpi, E.; Bely, B.; Bingley, M.; Britto, R.; Bursteinas, B.; Busiello, G.; et al. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47*, D506–D515.
36. Huang, Y.; Niu, B.F.; Gao, Y.; Fu, L.M.; Li, W.Z. CD-HIT Suite: A web server for clustering and comparing biological se-quences. *Bioinformatics* **2010**, *26*, 680–682. [CrossRef]
37. Madeira, F.; Park, Y.M.; Lee, J.; Buso, N.; Gur, T.; Madhusoodanan, N.; Basutkar, P.; Tivey, A.R.N.; Potter, S.C.; Finn, R.D.; et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **2019**, *47*, W636–W641. [CrossRef]
38. Stothard, P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA se-quences. *Biotechniques* **2000**, *28*, 1102–1104. [CrossRef] [PubMed]
39. *Open-Source PyMOL 1.6.0.0*; Schrodinger LLC: New York, NY, USA, 2013.
40. Zhang, Y.; Skolnick, J. TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **2005**, *33*, 2302–2309. [CrossRef]
41. Clausen, T.; Southan, C.; Ehrmann, M. The HtrA Family of Proteases: Implications for Protein Composition and Cell Fate. *Mol. Cell* **2002**, *10*, 443–455. [CrossRef]
42. Schubert, A.; Wrase, R.; Hilgenfeld, R.; Hansen, G. Structures of DegQ from Legionella pneumophila Define Distinct ON and OFF States. *J. Mol. Biol.* **2015**, *427*, 2840–2851. [CrossRef]
43. Teyra, J.; Ernst, A.; Singer, A.; Sicheri, F.; Sidhu, S.S. Comprehensive analysis of all evolutionary paths between two diver-gent PDZ domain specificities. *Protein Sci.* **2020**, *29*, 433–442. [CrossRef]
44. Jorda, J.; Xue, B.; Uversky, V.N.; Kajava, A.V. Protein tandem repeats—The more perfect, the less structured. *FEBS J.* **2010**, *277*, 2673–2682. [CrossRef] [PubMed]

45. Skelton, N.J.; Koehler, M.F.T.; Zobel, K.; Wong, W.L.; Yeh, S.; Pisabarro, M.T.; Yin, J.P.; Lasky, L.A.; Sidhu, S.S. Origins of PDZ domain ligand specificity—Structure determination and mutagenesis of the erbin PDZ domain. *J. Biol. Chem.* **2003**, *278*, 7645–7654. [CrossRef]
46. Krojer, T.; Sawa, J.; Schafer, E.; Saibil, H.R.; Ehrmann, M.; Clausen, T. Structural basis for the regulated protease and chap-erone function of DegP. *Nature* **2008**, *453*, 885-U31. [CrossRef] [PubMed]
47. Sun, R.; Fan, H.; Gao, F.; Lin, Y.; Zhang, L.; Gong, W.; Liu, L. Crystal Structure of Arabidopsis Deg2 Protein Reveals an Internal PDZ Ligand Locking the Hexameric Resting State. *J. Biol. Chem.* **2012**, *287*, 37564–37569. [CrossRef]
48. Burroughs, A.M.; Allen, K.N.; Dunaway-Mariano, D.; Aravind, L. Evolutionary genomics of the HAD superfamily: Understanding the structural adaptations and catalytic diversity in a superfamily of phosphoesterases and allied enzymes. *J. Mol. Bi-ol.* **2006**, *361*, 1003–1034. [CrossRef] [PubMed]
49. Vos, M.; Hesselman, M.C.; te Beek, T.A.; van Passel, M.W.J.; Eyre-Walker, A. Rates of Lateral Gene Transfer in Prokary-otes: High but Why? *Trends Microbiol.* **2015**, *23*, 598–605. [CrossRef]
50. Schaper, E.; Kajava, A.V.; Hauser, A.; Anisimova, M. Repeat or not repeat?—Statistical validation of tandem repeat predic-tion in genomic sequences. *Nucleic Acids Res.* **2012**, *40*, 10005–10017. [CrossRef] [PubMed]
51. Marold, J.D.; Kavran, J.M.; Bowman, G.D.; Barrick, D. A Naturally Occurring Repeat Protein with High Internal Sequence Identity Defines a New Class of TPR-like Proteins. *Structure* **2015**, *23*, 2055–2065. [CrossRef]
52. Gul, I.S.; Hulpiau, P.; Saeys, Y.; van Roy, F. Metazoan evolution of the armadillo repeat superfamily. *Cell. Mol. Life Sci.* **2016**, *74*, 525–541. [CrossRef] [PubMed]
53. Renault, L.; Nassar, N.; Vetter, I.; Becker, J.; Klebe, C.; Roth, M.; Wittinghofer, A. The 1.7 angstrom crystal structure of the regulator of chromosome condensation (RCC1) reveals a seven-bladed propeller. *Nature* **1998**, *392*, 97–101. [CrossRef] [PubMed]
54. D'Andrea, L.D.; Regan, L. TPR proteins: The versatile helix. *Trends Biochem. Sci.* **2003**, *28*, 655–662. [CrossRef] [PubMed]