Single-Domain Parvulins Constitute a Specific Marker for Recently Proposed Deep-Branching Archaeal Subgroups

Lederer, Christoph; Heider, Dominik; Boom, Johannes van den; Hoffmann, Daniel; Müller, Jonathan W.; Bayer, Peter

This text is provided by DuEPublico, the central repository of the University Duisburg-Essen. This version of the e-publication may differ from a potential published print or online version.

DOI: https://doi.org/10.4137/EBO.S7683
URN: urn:nbz:de:hbz:464-20180816-121315-8
Link: https://duepublico.uni-duisburg-essen.de:443/servlets/DocumentServlet?id=46772

Legal notice:
© the author(s), publisher and licensee Libertas Academica Ltd.
This is an open access article. Unrestricted non-commercial use is permitted provided the original work is properly cited. Creative Commons Attribution-NonCommercial 3.0 Unported License (CC BY-NC 3.0)

Source: Evolutionary Bioinformatics 2011:7 135–148; First Published September 19, 2011
Single-Domain Parvulins Constitute a Specific Marker for Recently Proposed Deep-Branching Archaeal Subgroups

Christoph Lederer¹, Dominik Heider², Johannes van den Boom¹, Daniel Hoffmann², Jonathan W. Mueller¹ and Peter Bayer¹
¹Departments for Structural and Medicinal Biochemistry, ²Bioinformatics, Centre for Medical Biotechnology (ZMB), Faculty of Biology, University of Duisburg-Essen, 45117 Essen, Germany.
Corresponding authors email: jonathanwmueller@web.de or jonathan.mueller@uni-due.de; peter.bayer@uni-due.de

Abstract: Peptidyl-prolyl cis/trans isomerases (PPIases) are enzymes assisting protein folding and protein quality control in organisms of all kingdoms of life. In contrast to the other sub-classes of PPIases, the cyclophilins and the FK-506 binding proteins, little was formerly known about the parvulin type of PPIase in Archaea. Recently, the first solution structure of an archaeal parvulin, the PinA protein from *Cenarchaeum symbiosum*, was reported. Investigation of occurrence and frequency of PPIase sequences in numerous archaeal genomes now revealed a strong tendency for thermophilic microorganisms to reduce the number of PPIases. Single-domain parvulins were mostly found in the genomes of recently proposed deep-branching archaeal subgroups, the Thaumarchaeota and the ARMANs (archaeal Richmond Mine acidophilic nanoorganisms). Hence, we used the parvulin sequence to reclassify available archaeal metagenomic contigs, thereby, adding new members to these subgroups. A combination of genomic background analysis and phylogenetic approaches of parvulin sequences suggested that the assigned sequences belong to at least two distinct groups of Thaumarchaeota. Finally, machine learning approaches were applied to identify amino acid residues that separate archaeal and bacterial parvulin proteins from each other. When mapped onto the recent PinA solution structure, most of these positions form a cluster at one site of the protein possibly indicating a different functionality of the two groups of parvulin proteins.

Keywords: archaeal protein, Pin1, PPIase, single-domain parvulin, Thaumarchaeota

*Evolutionary Bioinformatics* 2011:7 135–148
doi: 10.4137/EBO.S7683

This article is available from http://www.la-press.com.
© the author(s), publisher and licensee Libertas Academica Ltd.

This is an open access article. Unrestricted non-commercial use is permitted provided the original work is properly cited.
Introduction

Cis\textit{\slash}trans isomerisation of peptidyl-prolyl moieties within proteins can be regarded to be a molecular switch and is widely accredited as a means by which cell cycle events, protein (de)activation, folding and quality control are triggered. As cis\textit{\slash}trans isomerisation is a relatively slow process under moderate temperature, organisms of all three kingdoms of life have developed or maintained enzymes, the peptidyl-prolyl cis\textit{\slash}trans isomerases (PPIases), that accelerate this protein folding step.\textsuperscript{1,2} To date, three non-homologous families of PPIases are known: cyclophilins, FK506-binding proteins (FKBPs) and parvulins. Although cyclophilins and FKBPs have been analysed in several archaeal genomes and assigned chaperone activity in vitro, the actual cellular functions of these PPIases are not fully understood.\textsuperscript{1}

In contrast to cyclophilins and FKBPs, archaeal parvulin sequences only became available in the last few years by completion of the two genomes from \textit{Nitrosopumilus maritimus}\textsuperscript{3} and \textit{Cenarchaeum symbiosum},\textsuperscript{4} and by the deposition of three near-complete genomes of ultra-small acidophilic micro-archaea from biofilms from the Berkeley pit.\textsuperscript{5,6} In addition to these genomes, parvulin sequences are available from metagenomic studies using shotgun sequencing approaches on different samples: these were samples from fresh- and surface water\textsuperscript{7,8} and samples collected from iron mines.\textsuperscript{9} This last study described the archaeal Richmond Mine acidophilic nanoorganisms (ARMANs), acidophilic ultra small nano-archaea, which are frequent hosts of replicating viruses.\textsuperscript{10} These microorganisms seem to build a clade at the bottom of the euryarchaeal branch\textsuperscript{4} and are therefore annotated as “undefined Euryarchaeota” in NCBI databases.

The group of mesophilic Crenarchaeota was recently recognised as a new deep-branching phylum, the Thaumarchaeota.\textsuperscript{11} Recently, \textit{Nitrospheara gargensis} has been described as the first moderately thermophilic thaumarchaeal species based on its 16S ribosomal DNA sequence.\textsuperscript{12} There have been several studies that added few fosmids or metagenomic contigs to the phylum.\textsuperscript{13-15} However, the resolution within the thaumarchaeal cluster remained poor.

The cellular function of archaeal parvulins has not yet been studied in detail. Recently, the first structure of an archaeal parvulin, PinA of \textit{Cenarchaeaeum symbiosum}, a psychrophilic organism living in symbiosis with the marine sponge \textit{Axiella mexicana} has been solved.\textsuperscript{16} In the course of characterising structure and cellular function of archaeal single-domain parvulins (sdPars), we studied the occurrence and frequency of PinA proteins in different clades of the archaeal kingdom by a comparative genomics-based approach and defined by machine learning algorithms decisive structural features that separate bacterial and archaeal single-domain parvulins.

Results and Discussion

Distribution of prolyl isomerases in Archaea

We searched all available completely or nearly fully sequenced archaeal genomes for their content in prolyl isomerases (PPIases) to establish a basis for further phylogenetic analyses. A total of 98 genomes was investigated that can be grouped into 17 different groups including 13 established orders and the four recently proposed groups ARMANs,\textsuperscript{9} Thaumarchaeota,\textsuperscript{11} Korarchaeota\textsuperscript{17} and Nanoarchaeota.\textsuperscript{18} Their content in FK506 binding proteins (FKBPs), small and large cyclophilins, and parvulins is listed in Table 1; a detailed listing is given in Supplementary Table 1.

This comparison reveals that some PPIase subfamilies do not exist in Archaea: In contrast to bacterial and eukaryotic organisms, large multidomain FKBPs, multidomain cyclophilins (except Thaumarchaeota) and multidomain parvulins are absent in any archaeal genome sequenced so far. The only ubiquitous class of PPIases in Archaea are single-domain FKBPs except in \textit{Nanoarchaeum equitans} which is special because of its dependency on \textit{Ignicoccus}. While in all examined non-thaumarchaeal genomes, cyclophilins consist of a single PPIase domain of about 160 amino acids, the two known thaumarchaeal genomes of \textit{Nitrosopumilus maritimus} and \textit{Cenarchaeum symbiosum} additionally contain a protein with more than 500 amino acids carrying an N-terminal cyclophilin domain. No other conserved domains are found for the 545 aa \textit{C. symbiosum} protein in a CDD search.\textsuperscript{19} In the 509 aa \textit{N. maritimus} protein, there is a fragment of a putative Zn-dependent protease [CDD:COG4784] located in the middle of the protein. Although the function of this protein remains unclear, it separates the two thaumarchaeal species from the rest of the Archaea by its mere existence.
### Table 1. Occurrence of prolyl isomerases in archaeal genomes.

<table>
<thead>
<tr>
<th>Kingdom</th>
<th>Order</th>
<th>#</th>
<th>FKBPs</th>
<th>Cyclophilins</th>
<th>Single-domain parvulins</th>
<th>Temperature optimum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Single-domain</td>
<td>Multi-domain</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crenarchaeota</td>
<td>Desulfurococcales</td>
<td>8</td>
<td>1 per genome</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Thermoproteales</td>
<td>8</td>
<td>1 per genome</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Sulfolobales</td>
<td>12</td>
<td>1 per genome</td>
<td>1 in Picrophilus</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Archaeoglobales</td>
<td>3</td>
<td>1 per genome</td>
<td>1 in Ferroplasma</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Thermoplasmatales</td>
<td>4</td>
<td>1 per genome</td>
<td>1 in Thermoplasma</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Euryarchaeota</td>
<td>Thermococcales</td>
<td>9</td>
<td>1 in Pyrococcus</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2 in Thermococcus</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Halobacteriales</td>
<td>13</td>
<td>1–3 per genome</td>
<td>1 per genome</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Single-domain</td>
<td>Multi-domain</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Euryarchaeota</td>
<td>Methanococcales</td>
<td>12</td>
<td>2 in Methanocaldococcus; 2–3 in Methanococcus</td>
<td>None in Methanocaldococcus; 0–2 in Methanococcus</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Methanocecales</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Methanopyrales</td>
<td>1</td>
<td>1</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td></td>
<td>Methanosaicales</td>
<td>7*</td>
<td>1 per genome; 2 in Methanosaeta concilii</td>
<td>None</td>
<td>None</td>
<td>1 in Methanosaeta concilii</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Euryarchaeota</td>
<td>Methanobacteriales</td>
<td>8</td>
<td>1 per genome</td>
<td>1 per genome</td>
<td>None</td>
<td>none</td>
</tr>
<tr>
<td></td>
<td>Methanomicrobiales</td>
<td>6</td>
<td>1–4 per genome</td>
<td>1–2 per genome</td>
<td>None</td>
<td>1 in 4 genomes none in rest</td>
</tr>
<tr>
<td></td>
<td>ARMAN</td>
<td>3</td>
<td>2 in Micarchaeum; 3 in Pararchaeum</td>
<td>None</td>
<td>None</td>
<td>1 in two genomes none in 1 genome 1 per genome</td>
</tr>
<tr>
<td>Thaumarchaeota</td>
<td>–</td>
<td>3*</td>
<td>1 per genome</td>
<td>1 per genome</td>
<td>1 per genome; 2 in Methanosaeta concilii</td>
<td>None</td>
</tr>
<tr>
<td>Korarchaeota</td>
<td>–</td>
<td>1</td>
<td>1</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Nanoarchaeota</td>
<td>–</td>
<td>1</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Caldiarchaeum</td>
<td>–</td>
<td>1*</td>
<td>1, 2 or 3 (ambiguously annotated)</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
</tbody>
</table>

**Notes:** Refer to Supplementary Table 1 for details. *Please note that additional genome sequences became available very recently.*
From Table 1, a certain tendency for the content of PPIases can be inferred: The higher the preferred growth temperature, the lower the content of PPIase per genome. The same applies to the thermophilic *Korarchaeum cryptophillum* and the hyperthermophilic euryarchaeal orders Archaeoglobales and Methanopyrales. Strikingly, even different species from the same orders differ in their PPIase content depending on their favoured temperature range. The Methanococcales include the hyperthermophilic *Methanocaldococcus* and the mesophilic *Methanothermobacter* species. Whereas the hyperthermophilic species contain only two FKBP, mesophilic species additionally contain one or two cyclophilins or a third FKBP (e.g., *Methanothermobacter thermotrix*). Cold-adapted microorganisms contain more than three prolyl isomerases e.g., four PPIases are found in the psychrophilic archaean *Cenarchaeum symbiosum*.16

Although there are exceptions from the ‘rule’ when going to lower temperatures, the correlation itself is not surprising and was previously suggested.1 At higher growth temperatures spontaneous XaaPro bond isomerisation becomes faster and, hence, less enzymatic assistance in this process is needed. Although hyperthermophilic Archaea have reduced their PPIase repertoire to only one FKBP per genome, at least one isomerase seems to be absolutely crucial for them. Assuming nearly no difference between spontaneous and catalysed cis/trans isomerisation at elevated temperatures,1 this protein may serve a function other than being a PPIase.

In contrast to the total arsenal of PPIases, parvulin-type enzymes can only be found in three phyla including mesophilic microorganisms: the euryarchaeal Methanomicrobiales, the archaeal Richmond Mine acidophilic nanoorganisms (ARMAN) and the Thaumarchaeota. All Crenarchaeota, Korarchaeota and Nanoarchaeota species sequenced to date lack parvulin genes completely (Supplementary Table 1). Also very recent additions to the list of available archaeal genomes do not change this situation: The genomic sequence of another archaean (*Candidatus* Caldivirga subterraneum) that was recently released,21 contains a single-domain parvulin highly similar to the parvulin from *Nitrosporumilus maritimus* (85% identity on the level of amino acids). Thirdly, the genome of a Methanosarcina species (*Candidatus Methanoseta concilii* GP-6) containing a single-domain parvulin [NCBI RefSeq NC_015416.1] may indicate that the occurrence of parvulin coding sequences within the group of Euryarchaeota is not strictly confined to Methanomicrobiales.

Whereas only 9 percent of the 65 genomes of Euryarchaeota available at the time of analysis—including the above mentioned ARMAN and Methanomicrobiales—possess a parvulin gene, the two known thaumarchaeal genomes both contain exactly one parvulin gene. This parvulin comprises a single domain with a molecular weight of about 10 kDa. We refer to this class of parvulins as single-domain parvulins (sdPar). Single-domain parvulins are absent from eukaryotic genomes. It could be that the compartmentalised Eukarya need parvulin proteins with additional domains for cellular targeting, protein binding or anchoring like it is the case for the two human representatives Pin122,23 and Par14/17.24–27 In contrast to Eukarya, many bacterial genomes contain single-domain parvulins. With the exception of *Lentisphaera araneosa* and two Planctomycetes, all known parvulin-containing Bacteria belong to the subgroup of Proteobacteria. Most of these genomes possess a single sdPar; some species contain two, and only the extreme psychrophilic species *Colwellia psychrerythrocytaria* (Alteromonadales) contains three sdPar-type parvulin genes, which again suggests a relationship between temperature and PPIase content.

Of note, we found bacterial multi-domain proteins of the PrsA type containing parvulin domains very similar to sdPars. However, these paralogous sequences were excluded from further analysis because no corresponding multi-domain parvulin protein sequences could be found in any archaeal genome.

### Genomic context of archaeal parvulins

In order to characterise the relationship among archaeal parvulins, we examined the genomic context of the different parvulin loci in Archaea for conserved structures. In a first step, the genomic context of the parvulin locus was examined in the fully sequenced archaeal genomes. No conserved gene organisation was found within the
six available genomes from the Methanomicrobiales and the ARMAN group (Fig. 1, upper half). In contrast, the two deposited thaumarchaeal genomes of *Nitrosopumilus maritimus* and *Cenarchaeum symbiosum* have an antisense DEAD/DEAH-box helicase containing protein (DHC) downstream of the parvulin reading frame. The *Nitrosopumilus* genome carries an inserted hypothetical protein between parvulin and DHC, which is missing in the *Cenarchaeum* genome. In 5' direction from parvulin, both genomes contain two

![Diagram](image)

**Figure 1.** Genomic context analysis of the archaeal parvulin locus. The genetic background was analysed as described in the main text. White-backed arrows indicate genes occurring only once. Other colour codes are indicated within the figure. All abbreviations in this schematic are given below.

(Continued)
12 parvulin containing contigs starting with “ACY” from the Sorcerer II voyage,49 ACYX0108586 from samples collected from thick floating biofilms in the Richmond Mine.48 ABEF01053500 from a sub-tropical gyre in 4000 m depth was deposited by Ed DeLong and colleagues from the Hawaiian research station ALCHA. The parvulin containing fosmid AD1000-56-E4 derives from plankton collected in 1000 m depth in the Adriatic Sea.2 The metagenomic contigs ACYX022784421 and ACXJ0100858 were added to their group on basis of their parvulin primary sequence, for ACXJ0100858 this stands in agreement with its origin. The marine metagenomic contig ACYX023450473 contains 341 amino acids of the N-terminus of an amonofaceae class I/II. This sequence shows highest similarity to the protein YP_001737635 from Candidatus Korarchaeum cryptophyllum OPF3 (39 percent amino acids identity over 331 amino acids). Hence, this contig may belong to the Korarchaeota. ACYX02377022, AACYX00521263, ACXJ01072942 and AACYX01705999 were clustered with Nitrosopumilus maritimus, because sharing the same PPOX gene preceding parvulin. Contaminating AD1000-56-E4 was grouped with Candida archaebacal symbiosis due to an antisense hypothetical protein DUF2203 preceding the parvulin gene. With 80% sequence identity of their parvulin proteins, Nitrosopumilus maritimus and Candida archaebacal symbiosis were grouped together in Therm archaeota I. ACYX022114635, AACYX020921937. AACYX023014196 and AACYX025065072 were also clustered due to the gene directly upstream of parvulin, the hypothetical protein homologous to nmar_0940. AACYX021994642 was also grouped with these contigs because of its high parvulin primary sequence similarity. Thus, AD1000-56-E4 has a totally different upstream region, but they can clearly been classified as Therm archaeota I due to the typical downstream DHCP reading frame and their parvulin primary sequence.

eight of the metagenomic contigs the parvulin gene is found in close proximity to upstream genes (Fig. 1). In four contigs, a pyridoxamine 5′-phosphate oxidase-related FMN-binding protein (PPOX) is in direct neighbourhood to parvulin. Two of these four contigs also contain the same partial hypothetical protein 5′ to parvulin. This open reading frame is in direct vicinity to the parvulin locus in four other contigs, which lack the PPOX gene. Although the genome of *Nitrospopumilus maritimus* contains the reading frames for hypothetical protein nmar_0940, PPOX and sdPar (in this order), co-transcription is not likely there, because of an intergenic gap of more than 300 bases between parvulin and the PPOX gene. Hence, although the PPOX-gene could be co-transcribed with the parvulin in four contigs, this putative operon is most likely not conserved in all PPOX-containing organisms. Even though these findings give some hints for a polycistrionic transcription including the parvulin message, it is not possible to make functional statements as one of the two found proteins is a hypothetical protein of unknown function.

### Clustering of the highly conserved archaean single-domain parvulins

Besides the search for putative operons, a combination of genomic context analysis and comparison of the parvulin primary sequence suggested a grouping of the metagenomic contigs into different clusters indicated in Figure 1 that we refer to as Thaumarchaeota I and II. Genomes and contigs were added to the group of putative Thaumarchaeota I when either sharing a PPOX gene immediately upstream of the parvulin sequence (similar to the *N. maritimus* sequence) or the hypothetical protein DUF2203 (related to the *C. symbiosum* sequence). All other contigs comprising parvulin primary sequences homologous to *N. maritimus* or *C. symbiosum* were added to the second group of presumed Thaumarchaeota II. Nevertheless, all these sequences may belong to the formerly proposed group I.1a of Thaumarchaeota due to their overall similarity and their common oceanic origin.

Analysis of genomic context from fully sequenced archaean genomes and metagenomic data has now yielded archaean parvulin sequences in 23 different genomic environments. This allowed us to compare the diversity of sdPar amino acid sequences with that of other proteins. Therefore, we computed the sequence entropy, averaged over all sequence positions, for multiple sequence alignments of sdPars and multiple sequence alignments of proteins UbiA, DUF, hyp, PPOX, and DHCP from the corresponding organisms. Figure 2 shows that the mean sequence entropy of sdPars is significantly lower than that of other proteins from corresponding organisms (*P* = 0.01 in Mann-Whitney test). Thus, single-domain parvulins were found to be significantly more conserved than their genomic neighbours.

### Relationship among archaean parvulins

The high sequence conservation of sdPar proteins in Archaee tempted us to find out whether this short protein sequence (around 100 amino acids) allows the calculation of reasonable maximum likelihood phylogeny (MLP). Although parvulin cannot substitute for established phylogenetic markers like small subunit RNA, it can probably deliver valuable insights into the relationship within recently proposed archaean subgroups.

All available single-domain parvulins of Bacteria and Archaee were collected to generate an initial

![Figure 2. Mean Shannon entropy of the archaean parvulin and its genomic neighbours. At the parvulin locus of *N. maritimus*, we found the following neighbouring proteins to be present in at least 5 different contigs: UbiA, DUF, hyp, PPOX, and DHCP. The mean Shannon entropy (unit: bit) of these sequences was calculated as a measure of sequence diversity and compared with the same measure for sdPar from the corresponding organisms. Abbreviations: UbiA, UbiA prenyltransferase; DUF, hypothetical protein of unknown function (COG4911/DUF2203); hyp, hypothetical protein nmar_0940; PPOX, pyridoxamine 5′-phosphate oxidase-related FMN-binding protein; Par, parvulin; DHCP, DEAD/DEAH box containing protein.](image-url)
dataset of 326 sdPar sequences. Next, 59 bacterial paralogous sequences were removed, identified as such by calculating trees with the unfiltered dataset and comparing the results with a tree based on small subunit ribosomal RNA. Additionally, 26 parvulin sequences were filtered out from bacterial orders where less than 50 percent of fully or near fully sequenced genomes contained a single-domain parvulin that may have emerged from horizontal gene transfer (HGT). Trimming resulted in a final dataset of 241 sequences (218 bacterial genomes, eight archaeal genomes, one fosmid and 14 metagenomic contigs), which were aligned with T-Coffee and used to calculate maximum likelihood phylogeny with PhyML. Supplementary Figure 1 shows the resulting tree that was evaluated by 1000 bootstraps and rooted by *Escherichia coli* SlyD, an FKBP with structural analogy to the parvulin fold but without any sequential similarities.

Please note that the procedure described so far has assumed all single-domain parvulins to be monophyletic and under-estimates horizontal gene transfer (HGT) events. However, a recent paper has revealed that HGTs have been frequent events within marine uncultured planktonic archaea that Thaumarchaeota are part of. This paper also assigns a bacterial origin to the sdPar sequence of the archaeal fosmid AD1000-56-E4. Hence, we further concentrated on the relationships within the smaller archaeal sub-tree excluding all euryarchaeal sequences (Fig. 3). For this sub-tree to be monophyletic we have three main indications: (1) the respective sequences form a cluster set apart from other sdPar sequences by high bootstrap values (931/1000); (2) when using NmPinA or CsPinA sequences as seed for BLASTp searches within the nr database, the thaumarchaeal sdPars always constitute the very first hits with very small e values and (3) their genomic context is highly conserved.

Hence, it seems feasible to use single-domain parvulins as marker to detect and reclassify novel members of the recently proposed archaeal phyla of ARMANs and Thaumarchaeota. As one of the above mentioned metagenomic contigs (AACY023784421) does not show similarity to any of the ARMAN or thaumarchaeal sequences in its genomic context and its sdPar sequence clusters more with the four annotated Methanomicrobiales, this sequence may be of euryarchaeal origin.

---

**Figure 3.** Archaeal branch of an MLP tree combined with genomic context. The figure displays an expanded section of the maximum likelihood phylogeny tree from Figure 3. The outgroup has been omitted and the whole bacterial clade has been collapsed for clarity. Red signs indicate three deletion events suggested by the genomic context: The deletion of the hyp0940 and PPOX genes is described in the main text. The putative PPOX deletion seems to be a basal event for the Thaumarchaeota II subgroup. The large genetic rearrangement concerning the uncultured marine crenarchaeota fosmid AD1000-56-E4 makes this sequence unique in the group of Thaumarchaeota. Next to the Thaumarchaeota, the corresponding parvulin loci with the available genomic contexts are displayed. The groups predicted from the genomic context are also well defined in the MLP tree.
The parvulin sequence from the marine metagenomic contig AACY023450473 is different from all other sdPar sequences in the MLP tree. This contig contains a putative korarchaeal amino-transferase downstream of the sdPar reading frame (Fig. 1) and hence may be classified as a korarchaeal sequence. The respective microorganism may be the first mesophilic to moderately thermophilic korarchaeon as all archaeal parvulins found until now are from mesophilic species. Alternatively, it could be the first parvulin of a thermophilic species, as all Korarchaeota known are (hyper-)thermophilic. As this contig was from marine surface water samples, the first interpretation seems to be more likely. In either case, the existence of a parvulin within a korarchaeal genome supports that Korarchaeota may be genetic hybrids between Euryarchaeota and Crenarchaeota.

The sdPar sequences from all available ARMAN species are as a group highly similar to the respective proteins from Thaumarchaeota. This may indicate a closer relationship of ARMANs to Thaumarchaeota than to other Archaea. This would be in agreement with a recently reported 16s ribosomal RNA tree including the ARMANs where they branch very early from the euryarchaeal clade; but thaumarchaeal sequences were not enclosed in that tree. The ARMAN parvulins are similarly conserved as the thaumarchaeal within their group (score 69.7 and score 70.2, respectively), but unlike the Thaumarchaeota, they have no similarities to each other in their genomic context at all. Whether the ARMANs form another clade like Thaumarchaeota, Nanoarchaeota and Korarchaeota, or whether they belong to one of these clades remains to be elucidated. However, the comparison of the contained parvulin sequences makes a euryarchaeal annotation (“undefined Euryarchaeota”) of the group of extremely acidophilic organisms (ARMANs) rather unlikely.

The thaumarchaeal parvulin sequences are highly interrelated. This is in agreement with the recent acceptance of this group as a secluded deep branching phylum. Analysis of the genomic context suggests some major genetic rearrangements within this group. A comparison of the fully sequenced genomes of this group reveals a deletion of the hyp0940 and PPOX genes, present in *N. maritimus*, from the genome of the psychrophilic *C. symbiosum*.

Based solely on parvulin’s primary protein sequence, we were able to properly group the metagenomic contig AACY021994642 with only 804 bases, which is little more than half as long as the small ribosomal subunit RNA of *N. maritimus* (1409 bases). The other contigs of similar size, AACY023104196 (942 bases), AACY020912937 (797 bases) and AACY022114635 (989 bases), also contained information about the genomic context and could therefore be determined more reliably.

Our parvulin-based assignment of metagenomic contigs to distinct archaeal subgroups is of particular interest as it adds putative new members to the recently proposed phylum of Thaumarchaeota. Little is known today about the dimension, the diversity and the evolutionary relationships within this phylum, in spite of its important role in geochemical cycles in all marine surface waters on this planet. Using not only a single protein as a phylogenetic marker, but also its whole genomic locus including a variety of different and alternating genes gives additional opportunities of deepening the understanding of the phylum of Thaumarchaeota. One example how the toolset we deliver could be used in further studies is shown in Figure 4. Based on our analysis of the genomic context of parvulins, we propose sensible primers, which could be helpful for further studies. All proposed primers in neighbouring coding regions are within a 3 kb distance to the parvulin gene.

Comparing bacterial and archaeal parvulin proteins

Orthogonal to analyses of the PPIase repertoire within archaeal genomes, the genomic context and the degree of sequence conservation of archaeal parvulin proteins, we wanted to rationalise differences between these two groups of proteins on the level of amino acids. Qualitatively, protein parameters were compared for the parvulin proteins from *Escherichia coli* [PDB:1JNS] and *Cenarchaeum symbiosum* [PDB:2QRS]. The two proteins have similar isoelectric points: 9.23 and 9.59, respectively. However, the archaeal protein has more charged residues than the bacterial one (Asp+Glu/Arg+Lys: 12/18 relative to 9/13).

A more efficient and unbiased way to analyse differences between whole groups of protein sequences
Figure 4. Proposed primers for further metagenomic analyses. To get these primers the nucleotide sequences of the parvulin surrounding genes have been aligned with ClustalW. Several requirements (length between 18 and 24 bases, average GC content over 40%, average salt adjusted melting temperature between 50 °C and 65 °C) has been applied. The resulting primers for genes surrounding the thaumarchaeal parvulin are shown in this figure. For positions that were ambiguous, the respective IUPAC code for degenerate bases have been used: A or C, M; A or G, R; A or T, W; G or C, S; C or T, Y; G or T, K; A, G or C, V; A, C or T, H; A, G or T, D; G, C or T, B; A, G, C or T, N.

is to apply machine learning techniques.34 We used a random forest (RF)35 to differentiate our dataset of 241 archaean and bacterial parvulin sequences. First, the sequences were encoded using the hydrophobicity descriptor and the net charge descriptor, respectively (Fig. 5).36 The encoded sequences were used as input for the RF classifiers. The classifiers trained with the hydrophobicity and the net charge descriptors were perfectly able to distinguish the two classes. Remarkably, very good classification accuracy can already be achieved with a linear separator, such as a linear support vector machine (data not shown). For further analysis, we used the so-called importance values generated in the RF that project the complex classifier onto more easily intelligible contributions by single sequence positions.

Figure 5. A random forest (RF) can be trained to discriminate between bacterial and archaean single-domain parvulins. Two descriptors were used (hydrophobicity and net charge) to describe the protein sequences. The very same dataset used for the MLP tree was used here. Prediction scale represents “0.0”, bacterial, and “1.0”, archaean. The separation according to the MLP tree is represented in green (archaean) and red (bacterial). For the hydrophobicity descriptor, the RF perfectly separate the two classes (F1 score = 1.0). For the net charge descriptor, the RF reaches an F1 score of 0.979 (cut off = 0.2).
The positions were grouped into highly, medium or weakly important positions (Fig. 6A). A sequence logo was created from the 23 archaeal proteins for all these positions (Fig. 6B). It shows that the positions identified in the RF are largely conserved within the archaeal subgroup. A notable exception is Cys/Asp41 in the catalytic centre. Two of the 15 identified positions include charged amino acids in the archaeal proteins. This is in agreement with the notion above that an archaeal parvulin contains many more charged residues than its bacterial counterpart. Notably, the identified positions are unevenly distributed. After mapping of the corresponding residues onto the recently published structure of the PinA protein from *C. symbiosum*, these residues form a charged patch on one side of the protein (Fig. 6C). This may represent a special feature of archaeal parvulins pointing towards a negatively charged binding partner. As *C. symbiosum* is a psychrophilic archaeon, this feature could also be a further hallmark of cold adaption of psychrophilic parvulins.

**Conclusion**

Thaumarchaeota are known to play a crucial role in geochemical cycles in surface regions of non-coastal marine freshwater; their total number was estimated to be $10^{28}$ cells worldwide. These organisms do not belong to the phylum of Crenarchaeota, but form another deep-branching clade. Similar to Korarchaeota they possess genes related to Crenarchaeota as well as euryarchaeal genes. At the time of analysis, only two fully sequenced genomes of Thaumarchaeota are available and the dimension of this phylum is unknown. Recent studies have assigned some metagenomic contigs or fosmids to the thaumarchaeal phylum, but it remained impossible to group virtual organisms within the phylum.

Searching all available archaeal genomes for their PPIase content led to a correlation between the growth temperature of an organism and its PPIase content. There is a strong tendency for thermophilic microorganisms to reduce the total number of PPIases.

By examining the genomic context of sdPars, we could classify groups within the underrepresented and largely uncultivated archaeal subgroup of Thaumarchaeota and reveal that parvulin is significantly higher conserved than its genomic neighbours. Our work related the novel group of acidophilic Richmond Mine archaeal nanoorganism (ARMAN) much closer to the Thaumarchaeota than previously anticipated. Taken
together, our study significantly expands the phylum of Thaumarchaeota by metagenomic sequences, allows a first grouping of these organisms and reveals important amino acid residues, or a molecular phenotype, that separate archaeal and bacterial parvulins from each other.

**Methods**

**Searching genomes for their PPiase repertoire**

First all fully and near fully genomes of Archaea were listed using the NCBI genomes database. Next, the following *E. coli* proteins were used as queries: SlyD-type-FKBP: [GeneID: 947859]; FkpA-type-FKBP: [GeneID: 947870]; trigger factor: [GeneID: 945081]; Cyclophilins: [GeneID: 949038]; sdPar: [GeneID:948285]; SurA: [GeneID: 94481]. Queries for PrsA and NifM were taken from the organisms *Staphylococcus aureus* (PrsA:[GeneID: 5560626]) and *Azotobacter vinelandii* (NifM: [GeneID: 7759132]). The genomes were searched by examining their annotated proteins or by searching the whole genomes with the different queries using BLASTp. Positive, but not annotated hits were verified using the Conserved Domain Database (CDD).

**Retrieving sdPar-containing metagenomic data and measurement of diversity**

Metagenomic contigs were found by using the known and annotated archaeal parvulins as queries for tBLASTn searches in the whole genome shotgun (wgs) and the environmental samples (env) databases. The length of the sdPar reading frame was determined using ORF finder. Contigs were examined for the parvulin-surrounding genomic context using the same toolset.

For the evaluation of mean sequence entropies we computed with T-Coffee29 pairs of multiple sequence alignments for a certain protein sequence (UbiA, DUF, hyp, PPOX, DHCP) taken from a set of organisms and of sdPar taken from the same set. For each multiple sequence alignment we computed the Shannon entropy using the R-package bio3d41 and averaged this entropy over all alignment positions. The R-script for the computation of entropy and all multiple sequence alignments used as input for the script are provided as supplementary material (file Supp_entropy.zip).

**Phylogenetic calculations**

An initial dataset of 326 parvulin sequences was used for phylogenetic calculations. This dataset contained 16 sequences that were N-terminally truncated and one metagenomic sequence that has been C-terminally truncated to the sequence matching the sdPar sequences of *Nitrospumilus maritimus* and *Cenarchaeum symbiosum*. The FKBP SlyD from *E. coli* was used as an outgroup representing an unrelated prolyl isomerase with conserved fold. These sequences were aligned using T-Coffee. Phylogeny was estimated by maximum-likelihood using PhyML 3.030 with 1000 bootstraps. A consensus tree was derived from this dataset using the program consense of the Phylip suite defining SlyD manu-

**Machine learning approach**

Parvulins were first compared with respect to protein parameters and surface exposure as has been described. The filtered dataset of 241 single-domain parvulins was then used to train a linear support vector machine and a random forest. Therefore, N- and C-termini were trimmed on the basis of a multiple-sequence alignment and a loop of 15 amino acids was removed from sequences of *Photobacterium* species as it occurred only there. This S(E/Q)ALK(K/L)KNNLRGLI loop might functionally correspond to the phosphate-binding loop KHOSRRPPSWRQEKitRTRK of the Pin1 structure [PDB:1NMV];23 however, it is more similar to the KVKSKSKEGLD extension seen in the *Staphylococcus aureus* PrsA parvulin that does not bind phosphorylated substrates [PDB:2JZV]. The remaining sequences were projected to a length of 92 amino acids.

The 241 protein sequences were encoded using the hydrophobicity descriptor and the net charge descriptor, respectively. The encoded sequences were used as input for the linear support vector machine and the random forest as implemented in the R packages kernlab and randomForest. The classifier models were evaluated by ten-fold leave-one-out cross-validation. As performance measurement we used the area under
the receiver operating curve (AUC) and the F1 score, the harmonic mean of precision and recall:

\[
F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

\[
\text{precision} = \frac{TP}{TP + \text{FP}}, \quad \text{recall} = \frac{TP}{TP + \text{FN}}
\]

with TP: true positives, FP: false positives, FN: false negatives.

Random forests estimate the importance of each sequence position for the classification process. Importance values from random forests using hydrophobicity and net charge descriptors were averaged and classified into highly (>2%), medium (1%–2%) and weakly (0.6%–1%) important residues. These were mapped on the NMR structure [PDB:2QRS] of the sdPar of *C. Symbiosum*.

**Abbreviations**

aa, amino acid; ARMAN(s), archaeal Richmond Mine acidophilic nanoorganism(s); CDD, conserved domain database; contig, a contiguous stretch of aligned sequence; FKBP, FK506 binding protein, one class of PPIases; fosmid, DNA fragments in a vector based on the bacterial F-plasmid; HGT, horizontal gene transfer; MLP, maximum likelihood phylogeny; PPIase, (peptidyl-) prolyl isomerase; RF, random forest; sdPar, single-domain parvulin.

**Authors’ Contributions**

CL and JWM conceived this study. CL performed extensive database research. JvdB and DH performed MLP calculations. DH applied machine learning approaches. CL, JWM and PB analysed the data and wrote the paper. All authors read and approved the final manuscript.

**Acknowledgement**

We thank Bettina Siebers for constructive and critical reading of the manuscript. Die Brücke is acknowledged for fruitful discussions. This study was funded in part by DFG grant BA1624/7-1 to PB. Publication costs for this article were covered by the open access publishing program from Deutsche Forschungsgemeinschaft.

**Disclosures**

Author(s) have provided signed confirmations to the publisher of their compliance with all applicable legal and ethical obligations in respect to declaration of conflicts of interest, funding, authorship and contributorship, and compliance with ethical requirements in respect to treatment of human and animal test subjects. If this article contains identifiable human subject(s) author(s) were required to supply signed patient consent prior to publication. Author(s) have confirmed that the published article is unique and not under consideration nor published by any other publication and that they have consent to reproduce any copyrighted material. The peer reviewers declared no conflicts of interest.

**References**

Lederer et al


Publish with Libertas Academica and every scientist working in your field can read your article

“I would like to say that this is the most author-friendly editing process I have experienced in over 150 publications. Thank you most sincerely.”

“The communication between your staff and me has been terrific. Whenever progress is made with the manuscript, I receive notice. Quite honestly, I’ve never had such complete communication with a journal.”

“LA is different, and hopefully represents a kind of scientific publication machinery that removes the hurdles from free flow of scientific thought.”

Your paper will be:

- Available to your entire community free of charge
- Fairly and quickly peer reviewed
- Yours! You retain copyright

http://www.la-press.com