Archaea 2, 127–135

The genome of *Hyperthermus butylicus*: a sulfur-reducing, peptide fermenting, neutrophilic Crenarchaeote growing up to 108 °C

KIM BRÜGGER,[1,2] LANMING CHEN,[1,2] MARKUS STARK,[3,4] ARNE ZIBAT,[4] PETER REDDER,[1] ANDREAS RUEPP,[4,5] MARIANA AWAYEZ,[1] QUNXIN SHE,[1] ROGER A. GARRETT[1,6] and HANS-PETER KLENK[3,4,7]

[1] *Danish Archaea Centre, Institute of Molecular Biology, Copenhagen University, Sølvgade 83H, 1307 Copenhagen K, Denmark*

[2] *These authors contributed equally to the project*

[3] *e.gene Biotechnologie GmbH, Poeckinger Fussweg 7a, 82340 Feldafing, Germany*

[4] *Formerly EPIDAUROS Biotechnologie AG, Genes and Genome Analysis Team*

[5] *Present address: Institut für Bioinformatik, GSF-Forschungszentrum für Umwelt und Gesundheit, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany*

[6] *Editing author*

[7] *Corresponding author (hans-peter.klenk@online.de)*

Received October 26, 2006; accepted January 2, 2007; published online January 19, 2007

Summary *Hyperthermus butylicus*, a hyperthermophilic neutrophile and anaerobe, is a member of the archaeal kingdom Crenarchaeota. Its genome consists of a single circular chromosome of 1,667,163 bp with a 53.7% G+C content. A total of 1672 genes were annotated, of which 1602 are protein-coding, and up to a third are specific to *H. butylicus*. In contrast to some other crenarchaeal genomes, a high level of GUG and UUG start codons are predicted. Two *cdc6* genes are present, but neither could be linked unambiguously to an origin of replication. Many of the predicted metabolic gene products are associated with the fermentation of peptide mixtures including several peptidases with diverse specificities, and there are many encoded transporters. Most of the sulfur-reducing enzymes, hydrogenases and electron-transfer proteins were identified which are associated with energy production by reducing sulfur to H_2S. Two large clusters of regularly interspaced repeats (CRISPRs) are present, one of which is associated with a crenarchaeal-type *cas* gene superoperon; none of the spacer sequences yielded good sequence matches with known archaeal chromosomal elements. The genome carries no detectable transposable or integrated elements, no inteins, and introns are exclusive to tRNA genes. This suggests that the genome structure is quite stable, possibly reflecting a constant, and relatively uncompetitive, natural environment.

Keywords: anaerobe, genome analysis, hyperthermophile, solfataric habitat.

Introduction

Hyperthermus butylicus was isolated from a solfataric habitat with temperatures of up to 112 °C that lies on the sea-bed off the coast of the island of São Miguel in the Azores (Zillig et al. 1990). It grows between 80 and 108 °C with a broad temperature optimum. The organism utilizes peptide mixtures as carbon and energy sources but not amino acid mixtures, various synthetic peptides or undigested protein. It can also generate energy by reduction of elemental sulfur to yield H_2S. Fermentation products include CO_2, 1-butanol, acetic acid, phenylacetic acid and a trace of hydroxyphenyl acetic acid, which are produced in the presence or absence of sulfur and hydrogen (Zillig et al. 1990).

To date, few crenarchaeal genomes have been completely sequenced and analyzed. Those that are published are limited to three *Sulfolobus* species (family *Sulfolobaceae,* order Sulfolobales), *A. pernix* (family *Desulfurococcaceae,* order Desulfurococcales) and *P. aerophilum* (family *Thermoprotaceae,* order Thermoproteales) (reviewed in Klenk 2006). Thus, our genome-based knowledge of crenarchaeal biology is biased to the *Sulfolobaceae*. Moreover, this bias is reinforced by the fact that most of the characterized crenarchaeal extrachromosomal elements, including many cryptic and conjugative plasmids, as well as several novel viruses, are associated with this family (Zillig et al. 1998, Greve et al. 2004, Prangishvili and Garrett 2005). The *H. butylicus* genome presented here is the first to be reported for the family *Pyrodictiaceae* (order Desulfurococcales) and constitutes an important step in extending our knowledge of crenarchaeal diversity and of thermal biology because no other genome sequence is available for an organism with such a high optimal growth temperature.

Materials and methods

Genome sequencing

The DNA for genome sequencing of *H. butylicus* type strain

DSMZ 5456 (NCBI Taxonomy ID 54248) was taken from the original preparation (Zillig et al. 1990, 1991). The genome was cloned and mapped using a shot-gun strategy with plasmid vector pUC18 (average insert fragment size 2.5 kbp), with no additional large insert library. Template preparation for sequencing was performed with Qiagen robots (Qiagen Westburg, Germany) and ABI (Applied Biosystems, Foster City, USA) and MegaBACE 1000 Sequenators (Amersham Biotech, Amersham, U.K.). The genome was assembled from 9313 sequence reads with a mean trimmed read length of 739 nt, resulting in 4.4-fold mean sequence coverage. For gap closure and sequence editing, about 200 primer walking reactions were performed on plasmid clones. Several sequence ambiguities were resolved by generating and sequencing appropriate PCR fragments. The genome was assembled with the Phred-Phrap-Consed software package (Ewing and Green 1998, Ewing et al. 1998, Gordon et al. 1998).

Gene identification and functional annotation

Whole genome annotation and analysis was performed in the MUTAGEN system (Brügger et al. 2003). Protein coding genes were identified with the bacterial gene finder EasyGene (Larsen and Krogh 2003). Functional assignments are based on searches against GenBank (Benson et al. 2003), COG (Tatusov et al. 2001), SWISS-PROT (Boeckmann et al. 2003) and Pfam databases (Bateman et al. 2002). Transmembrane helices where predicted with TMHMM (Krogh et al. 2001) and signal peptides with SignalP (Bendtsen et al. 2004). tRNA genes were located with tRNAscan-SE (Lowe and Eddy 1997). Open reading frames (ORFs) with homologs in at least two other genomes were inferred to constitute genes and are included in the final annotation. After a round of manual annotation, remaining frameshifts were checked by generating and sequencing PCR products. Remaining frameshifts were considered to be authentic. Finally, the annotations were checked by a different annotator to validate their correctness.

Results and discussion

Genome sequencing and annotation

The 1,667,163 bp circular genome of *H. butylicus* was analyzed and annotated using MUTAGEN (Brügger et al. 2003). About 1672 genes were identified, of which 1602 are protein-coding; only one of these, a *CBf5* gene (Yoshinari et al. 2006), has been published previously. The number of predicted protein genes that are homologous with the other sequenced archaeal hyperthermophilic neutrophiles, *A. pernix* (Kawarabayasi et al. 1999) and *P. aerophilum* (Fitz-Gibbon et al. 2002), are illustrated (Figure 1A). These data demonstrate that each of the analyzed genomes carries a large number of genes, exclusive to that genome, in the range 41–44% for *H. butylicus* and *A. pernix*, and 62% for the larger genome of *P. aerophilum*, with a relatively low proportion of the genes (40, 36 and 26%, respectively) shared among all three organisms. These results serve to underline the considerable phylogenetic diversity among hyperthermophilic crenarchaea. The second plot (Figure 2B) demonstrates a much lower degree of gene homology between *H. butylicus* and the euryarchaeal hyperthermophile *M. jannaschii* (16%) and the bacterial hyperthermophile *Aquifex aeolicus* (6%), which extend over much larger evolutionary distances.

Although most archaeal genes are predicted to use an AUG start codon, for *H. butylicus* a large percentage of the predicted start codons were GUG (25%) or UUG (37%) using the Easygene prediction program (Larsen and Krogh 2003, Torarinsson et al. 2005). Moreover, similar values were obtained for the archaeal hyperthermophiles *A. pernix* and *Methanopyrum kandleri*, whereas *P. aerophilum* was predicted to employ 32%

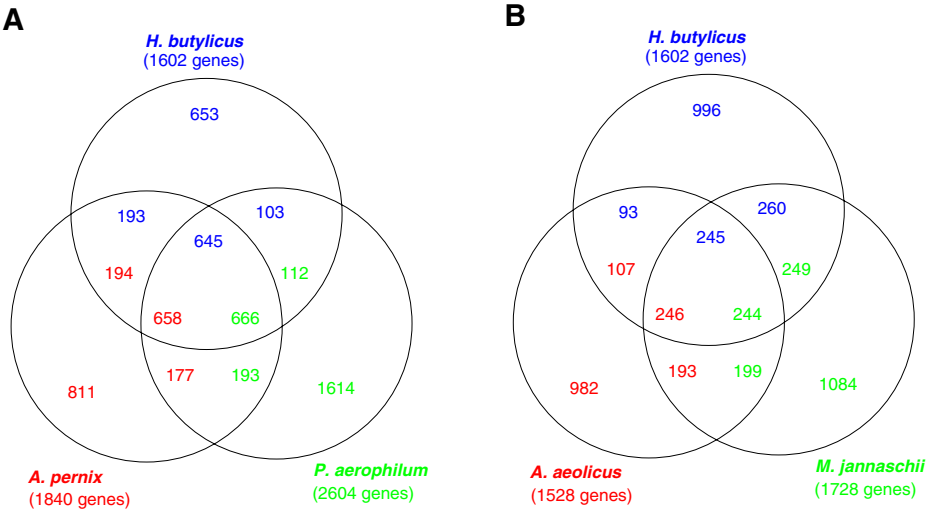

Figure 1. Homologous protein genes of *H. butylicus* and other hyperthermophiles. The overlapping circle plot shows the number of homologs shared between the genomes of A. *H. butylicus*, *A. pernix* and *P. aerophilum*, chosen as representatives of the three major crenarchaeal neutrophile families for which genome sequences are available, and B. *H. butylicus*, *M. jannaschii* and *A. aeolicus* representing hyperthermophilic crenarchaea, euryarchaea and bacteria, respectively. Blue, red and green numbers each show the number of genes of one genome which are specific to that genome or which are shared (homologous) with one, or both, of the other two genomes.

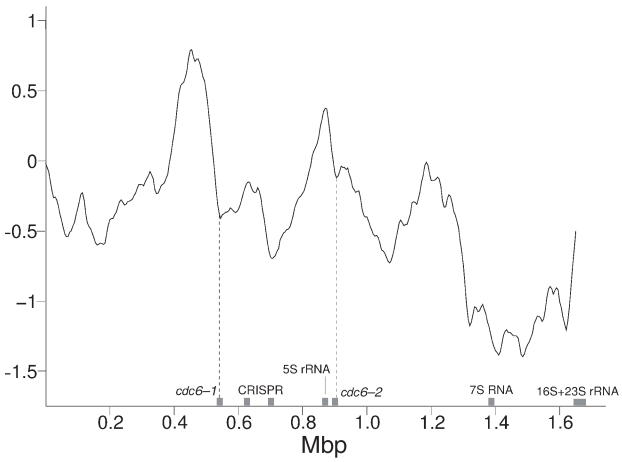

Figure 2. A GC-skew analysis of the *H. butylicus* genome showing the major peaks where nucleotide compositional deviations occur. Shown below are some of the main features of the genome including the two *cdc6* genes, rRNA genes and large repeat clusters (CRISPR). The ordinate scale is arbitrary, showing positive and negative G/C deviations. The abscissa corresponds to the genome sequence length where the numbering starts just upstream from the 16S + 23S rRNA operon.

GUG but few UUG start codons (Torarinsson et al. 2005).

To test the reliability of these conclusions, we examined a selection of the *H. butylicus* protein coding genes predicted to use non AUG start codons. Of the 1602 identified genes, 981 were predicted to start with GUG or UUG codons. These were aligned with any conserved homologous genes from other species of crenarchaea or archaea predicted to use AUG as the primary start codon. A total of 191 genes yielded multiple alignments, where the homologous genes started within ± 45 bp of the start of the *H. butylicus* gene. Of these 191 *H. butylicus* genes, 42 genes exhibited possible alternative AUG start codons within ± 45 bp of the predicted start site, and 149 genes did not. These results strongly reinforce that the usage of GUG and UUG start codons is widespread in *H. butylicus*.

Twelve genes, which are conserved in other genomes, exhibit potential frameshifts in *H. butylicus*. Two of these carry "slippery" sequences, one of 5 As (Hbut_1481) and another of 6 Cs (Hbut_1362), which could facilitate the coding of functional proteins. Functional frame-shifts have also been reported for *S. solfataricus* P2 (Watanabe et al. 2002, Cobucci-Ponzano et al. 2003). The values were much lower than the 32 genes identified in *P. aerophilum* that were predicted to carry one or more frameshifts. The latter were assumed to occur mainly at homopolynucleotide sequences, and the high frequency was attributed to the lack of an efficient mismatch repair system (Fitz-Gibbon et al. 2002).

The G+C content of 53.7% is close to the original estimate (Zillig et al. 1990) and consistent with genomic values obtained for other sequenced archaeal hyperthermophilic neutrophiles, and is substantially higher than values determined for hyperthermophilic acidophiles (32–37% G+C). The high A+T values obtained for the latter may reflect the relative instability of purines and cytosines at low pH if the DNA is sometimes exposed to the extracellular environment (Lin and Forsdyke 2007).

Surfaces of hyperthermophile proteins are estimated to carry a relatively high proportion of the charged residues glutamic acid, arginine and lysine and a relatively low proportion of the non-charged polar residues, mainly glutamine (Cambillau and Claverie, 2000). This pattern is consistent with the annotated genome of *H. butylicus*, where the percentage of codons for the charged amino acids is about 20%, compared with an average of 16% for mesophilic archaeal genomes, whereas the percentage of predicted glutamine codons at 1.9% is about half that found in the mesophilic genomes.

The genome is numbered from just upstream of the 16S + 23S rRNA genes, and the genome sequence is deposited in EMBL/GenBank under Accession Number CP000493.

Genome stability

No evidence was found for active, transposable, insertion sequence (IS) elements. However, two transposase gene fragments were detected, Hbut_0323 and Hbut_1361, where the latter lacks the gene start and is interrupted by a UAG stop codon. Both fragments are homologous to ISC1314 which occurs in two copies in *A. pernix* (Brügger et al. 2002). A search for imperfect repeat sequences using the program LUNA (K.B. unpublished) revealed no miniature inverted repeat transposable elements (MITEs), which occur in some crenarchaeal genomes where IS elements are present (Redder et al. 2001, Brügger et al. 2002). Furthermore, searches against the Sulfolobus Database (Brügger 2007) yielded no evidence for the presence of integrated crenarchaeal viruses, plasmids or other genetic elements. Nor were any *att* sites detected, which could have indicated earlier integration activity (She et al. 2004), although a single integrase gene is present (Hbut_1547). Thus, the genome appears to have been minimally influenced, at least in recent times, by either mobile element transposition or integration, and this inference is consistent with a fairly even G-C base pair distribution throughout the genome (data not shown).

Carbon metabolism

Hyperthermus butylicus is an anaerobe and carries two genes for detoxification of O_2, a superoxide reductase (Hbut_1161) and a peroxyredoxin (Hbut_0228), which remove superoxide without producing O_2, thus maintaining a reduced state within the cell.

Consistent with the capacity to ferment peptides (Zillig et al. 1990), the *H. butylicus* genome encodes at least 23 putative peptidases that belong to diverse families and include peptidases, oligopeptidases, metallopeptidases and amino- and carboxy-peptidases as well as more specific enzymes such as a pyroglutamyl peptidase (Hbut_1614) and a glycoprotease (Hbut_0434). At least five of these carry putative N-terminal signal peptides suggesting that they are active extracellularly.

When supplied with tryptone, *H. butylicus* produces a number of by-products including CO_2, 1-butanol, acetate, propionate, phenylacetate, hydroxyphenylacetate, acetophenone, hy-

droxyacetophenone and propylbenzene (Zillig et al. 1990). Some of these are likely to be products of amino acid degradation. For example, propionate can be produced from propionyl-CoA, a degradation product of valine, leucine and isoleucine, possibly by Hbut_0129, Hbut_0995 or Hbut_1337. However, genes encoding some enzymes from this degradation pathway, including enoyl hydratase, acyl-CoA dehydrogenase and a branched-chain amino acid aminotransferase, were not detected, suggesting that an alternative pathway exists in *H. butylicus*.

Phenylacetate is likely to be produced from phenylalanine via 2-phenylacetamid (Hbut_0228 and Hbut_0594), but since no mechanism is known for converting the fermentation product phenylacetate to hydroxyphenylacetate, the latter probably arises from tyrosine degradation, either via the pathway described above or by converting tyrosine to tyramine (Hbut_0224) and then to hydroxyphenylacetate by two reactions for which the relevant genes were not yet detected. The latter product probably cannot be metabolized further because *H. butylicus* lacks 4-hydroxyphenylacetate-3-hydroxylase, 4-hydroxyphenylacetate 1-hydroxylase and phenylacetate-CoA ligase, which occur in some other species of archaea.

Although 1-butanol is produced, no homolog of the bacterial acetoacetate-butyrate/acetate coenzyme A transferase was found, which is essential for 1-butanol production in *Clostridium* species (Toth et al. 1999). Acetate, another major by-product, is involved in cysteine biosynthesis where *O*-acetyl-L-serine + H_2S yield L-cysteine and acetic acid may be catalyzed by Hbut_0113 or Hbut_1585. However, AMP can also be converted to ATP by splitting acetyl-CoA to CoA and acetic acid.

It is unclear whether *H. butylicus* produces 2-hydroxyacetophenone, 4′-hydroxyacetophenone or both (Zillig et al. 1990). Acetophenone and 2-hydroxyacetophenone are products of ethylbenzene degradation, with the latter produced from the former. A range of enzymes are candidates for the production of acetophenone from 1-phenylethanol. 4′-hydroxyacetophenone, however, may be produced from 1-(4′-hydroxyphenyl) ethanol.

The genome encodes a wide range of transporter proteins. Thirty-seven genes encode predicted ABC-type transporters. Moreover, five gene products were implicated in cation transport and three in anion transport. In addition, six genes encoded proteins implicated in branched chain amino acid transport.

We infer that the cell appendages implicated in cell motility (Zillig et al. 1990) were not flagella because homologs of genes encoding archaeal flagellin (FlaB), crucial for flagella formation, as well as other flagella-associated genes, are absent (Ng et al. 2006, S. Gribaldo, Institute Pasteur, Paris, pers. comm.).

Sulfur metabolism

Some archaeal enzymes involved in metabolizing and transporting sulfur products have been characterized experimentally. In particular, functional protein complexes from both *Pyrodictium abyssi* (Dirmeier et al. 1998) and *Acidianus ambivalens* (Laska et al. 2003), which are membrane-bound and contain multiple subunits, were shown to catalyze sulfur reduction to H_2S. The protein complexes are associated with sulfur reductase, hydrogenase and electron transfer activities, and, at least for *A. ambivalens*, the relevant genes are clustered in operons. Sulfur reduction to H_2S is an important energy generating process in *H. butylicus* (Zillig et al. 1990), and homologs of many of these characterized genes are clustered in operons in the genome. They include those for sulfur-reducing enzymes *sreA* (Hbut_0373, Hbut_1051), *sreB* (Hbut_0372, Hbut_1052) and *sreD* (Hbut_0757) and an operon containing *sdhE* (Hbut_0155), *sdhF* (Hbut_0156) and *sdhB* (Hbut_0157). Genes encoding hydrogenases, and other proteins considered necessary for the function of the sulfur-reducing complex, include *hypC* (Hbut_0754), *hypD* (Hbut_0755) and an operon carrying *hydL* (Hbut_1368), *isp2* (Hbut_1369), *isp1* (Hbut_1370), *hydS* (Hbut_1371) and *hoxM* (Hbut_1372), whereas *tatC* (Hbut_1508) and *tatA* (Hbut_1509) are implicated in transport (Laska et al. 2003, Kletzin 2006).

DNA replication

Archaea often carry multiple *cdc6* genes, and for *Pyrococcus* and *Sulfolobus* species, some of these have been shown experimentally to be located close to chromosomal replication origins (Matsunaga et al. 2001, Lundgren et al. 2004, Robinson et al. 2004). *Hyperthermus butylicus* has two genes, *cdc6-1* (Hbut_0595) and *cdc6-2* (Hbut_0909), where the former is preceded by 1700 bp lacking ORFs, whereas the latter lies within an operon. Generation of Z-curve and G+C-skew plots (Zhang and Zhang 2005) yielded a few peaks and troughs that could represent replication origin(s). Although *cdc6-2* gives the best match at around position 0.9 Mb in the G+C skew plot, neither gene coincides exactly with a peak or trough (Figure 2). Moreover, no ORB motifs characteristic of archaeal chromosomal origins (Robinson et al. 2004) were detected adjoining the *cdc6* genes (within 2 kb on either side), although this could reflect a low level of sequence conservation of the motifs with those of *A. pernix* and other crenarchaea. The normal complement of genes associated with the crenarchaeal replication apparatus were present (Marsh and Bell 2006). Two chromatin/RNA binding Alba homologs are encoded (Hbut_0909 and Hbut_0977).

DNA repair and modification

The genome contains many of the previously characterized archaeal DNA repair enzymes, including some involved in double-strand-break repair, direct repair, base excision repair and nucleotide excision repair (White 2006). For direct repair, 6-*O*-methylguanine DNA methyltransferase is encoded, and for the base excision pathway, several proteins are encoded. Moreover, consistent with the sea bed environment from which *H. butylicus* was isolated (Zillig et al. 1990), and in con-

trast to some other species of archaea (e.g. Baliga et al. 2004, Chen et al. 2005), no genes were detected for extra UV excision repair pathways. Moreover, no gene was found for a deoxyribodipyrimidine photolyase (COG0415) which could directly repair pyrimidine dimers generated by visible light, nor for 8-oxoguanine DNA glycosylase, which repairs the lesion of 8-oxoguanine (*oxoG*) caused by ionizing radiation or oxidizing agents (Gogos and Clarke 1999).

Furthermore, no *H. butylicus* gene was detected for an error-prone DNA polymerase which can accommodate a variety of modified templates and achieves trans-lesion synthesis by different mechanisms (Yang 2005). As for most sequenced archaeal genomes, genes encoding the eukaryal/bacterial-type mismatch repair proteins were not observed in the *H. butylicus* genome. Although a common DNA repair photolyase (Hbut_1189, COG 1533) and two RecB-like exonucleases (Hbut_0039, Hbut_0849) are encoded, their precise roles in DNA repair remain to be determined.

Transcription and translation

The total complement of crenarchaeal RNA polymerase subunits is encoded (Zillig et al. 1990), with a single *rpoB* gene as occurs in almost all crenarchaeal genomes (She et al. 2001). Genes for the characteristic archaeal transcription factors are also present, with a single gene for the TBP protein (Hbut_0934) and two genes for TFIIB (Hbut_0054 and Hbut_0458). For the translational apparatus, single copies of the coupled 16S and 23S rRNA genes and a single, uncoupled, 5S rRNA gene are present as well as a single 7S RNA gene and the usual complement of crenarchaeal ribosomal protein genes (26 small subunit and 38 large subunit). The ribosomal factors involved in initiation, elongation and termination of protein synthesis are also typically crenarchaeal (Londei 2006).

Most genes (74%) are preceded by Shine-Dalgarno (S-D) sequence motifs and, for many of these, predicted TATA-like motifs occur at least 25 bp further upstream. Thus, *H. butylicus* is unlikely to produce large numbers of leaderless transcripts as was predicted to occur for some other crenarchaea (Torarinsson et al. 2005). As shown earlier, archaeal S-D motifs are more variable than those of bacteria, and their sequences depend on the actual length of the conserved 3′-terminal sequence of 16S rRNA (Torarinsson et al. 2005). Exceptionally, *H. butylicus*, together with *A. pernix* and *M. kandleri*, shows a high incidence of predicted GGGG S-D motifs (Torarinsson et al. 2005).

The genome of *H. butylicus* contains 46 tRNA genes carrying 43 different anti-codons coding for 20 amino acids. As for most archaeal genomes, three genes are present for $tRNA^{Met}$(CAT), which differ in sequence with one containing an intron located at the anti-codon loop. One of these is classified as an initiator tRNA (Hbut_1108) on the basis of sequence comparison with an experimentally characterized archaeal initiator tRNA (Kuchino et al. 1982). As for other crenarchaeal genomes, a selenocysteine incorporation system is lacking, but a selenocycteine lyase (Hbut_0769) is encoded for salvaging cysteine.

Genes for several putative RNA modifying enzymes were identified, consistent with a relatively large modification of the archaeal hyperthermophile RNAs (Rozenski et al. 1999). Searches for other untranslated RNAs, using experimentally determined sequences from other crenarchaea, failed to reveal unambiguous RNA homologs in *H. butylicus*. We infer, therefore, that sequence conservation is too low and that these RNAs should be identified experimentally.

Cellular defence systems

Although no unambiguous gene matches were found for restriction enzymes, which may reflect the limited sequence conservation of this enzyme family, a putative DNA-modifying C-5 cytosine methylase is encoded (Hbut_1135). Moreover, *H. butylicus* contains two large clusters of short regular spaced repeats (SRSR/CRISPR) positioned 66.5 kb apart and containing 46 and 47 repeat-spacer units (Figure 2), and encodes a homolog of the protein (Hbut_0986) that interacts specifically with the repeat sequence in the *S. solfataricus* P2 chromosome and *Sulfolobus* plasmid pNOB8 (Peng et al. 2003). The spacer sequences are considered to derive from extrachromosomal elements and, in particular, from viruses (Mojica et al. 2005), and their spacer transcripts (Tang et al. 2002, 2005) may inactivate viral or plasmid propagation by acting at the level of mRNA or DNA (Lillestøl et al. 2006, Makarova et al. 2006). No matches were observed between the spacer sequences of *H. butylicus* and any known archaeal viruses and plasmids, but this negative result may reflect the unavailability of extrachromosomal genetic elements which can propagate in *H. butylicus* or in related organisms.

A single superoperon containing a set of *cas* and *csa* genes adjoins one of the repeat clusters. The encoded proteins are considered to be involved in the development and activity of the repeat-clusters (Jansen et al. 2002, Lillestøl et al. 2006, Makarova et al. 2006). The gene order is crenarchaea-specific (Lillestøl et al. 2006), except that the *cas6* gene, which is generally located most distantly from the repeat-cluster, is absent. These genes are predicted to encode diverse nucleases *cas1* (Hbut_0652), *cas4* (Hbut_0649), *csa2* (Hbut_0644), a DNA helicase *cas3* (Hbut_0640) and an RNA binding protein *cas5* (Hbut_0643) (Makarova et al. 2002, Haft et al. 2005, Makarova et al. 2006).

Given the apparent stability of the *H. butylicus* genome, which may be partly due to an efficient restriction-modification system, as has been argued for the similarly stable genome of *S. acidocaldarius* (Chen et al. 2005), the large number of repeat-spacer units in the clusters is difficult to explain. The current hypothesis for the clusters (Mojica et al. 2005) suggests that the organism has been regularly invaded by extrachromosomal elements. If true, then they have, surprisingly, left no detectable traces in the genome other than the spacers. One possible explanation is that the large archaeal clusters are multifunctional and have other important activities which the smaller, non ubiquitous, bacterial clusters lack.

Introns and inteins

Introns occur within 10 of the 46 tRNA genes. They are all archaeal-type introns and are located +1 bp after the 3′-end of the anticodon, as occurs for many archaeal tRNA introns and for all eukaryal tRNA introns (Marck and Grosjean 2003). The origin of the archaeal tRNA introns is still unclear. In archaea, rRNA intron mobility is generally facilitated by a homing endonuclease (Aagaard et al. 1995), but none appears to be encoded in the *H. butylicus* genome. Another hypothesis is that they arise by fusion of fragmented tRNA genes (Randau et al. 2005). A comparison of the tRNA-containing introns of *H. butylicus, A. pernix* and *P. aerophilum* (Table 1) reveals some conservation of tRNA-containing introns between *H. butylicus* and *A. pernix* (although there is no general conservation of sequence) but only three conserved tRNAs between these two organisms and *P. aerophilum* (Met (CAT), Pro (GGG) and Tyr (GTA)). On balance, it suggests, as was concluded from a study of the three *Sulfolobus* genomes (Brügger et al. 2006), that the introns can be mobilized, but presumably only when a homing endonuclease is present as it is in *P. aerophilum* (Burggraf et al. 1993).

No introns are present in the rRNA genes although they occur widely in other crenarchaea. To date, only one archaeal intron has been demonstrated to occur within a protein coding sequence (Watanabe et al. 2002) although others have been predicted (Brügger et al. 2006). The identified intron is located in homologs of the eukaryal Cbf5 protein gene which is a subunit of the small nucleolar RNA-protein complex. Its excision from mRNAs in *A. pernix*, *S. solfataricus* and *S. tokodaii* removes a frameshift from the coding region (Watanabe et al. 2002). However, the *H. butylicus Cbf5* gene (Hbut_1329) lacks this intron, and this correlates with the recent study showing that the introns are confined to all studied members of the crenarchaeal families *Sulfolobaceae* and *Desulfurococcaceae* (Yoshinari et al. 2006).

Two components of the intron splicing enzyme are encoded, a catalytic subunit (Hbut_0054) and a structural subunit (Hbut_0955), which together generate the oligomeric structure for the archaeal intron-splicing enzyme characterized for

Table 1. Comparison of intron-containing tRNAs.

tRNA (anti-codon)	*H. butylicus*	*A. pernix*	*P. aerophilum*
Ala (TGC)			+
Arg (TCT)		+	
Asn (GTT)			+
Asp (GTC)		+	
Cys (GCA)	+	+	
Gln (CTG)			+
Gln (TTG)	+		
Gly (CCC)			+
Gly (TCC)			+
His (GTG)			+
Lys (CTT)	+	+	
Lys (TTT)		+	
Met (CAT)	+	+/+	+
Pro (CGG)		+	
Pro (GGG)	+	+	+
Pro (TGG)			+
Ser (CGA)	+	+	
Ser (GCT)			+
Thr (CGT)	+	+	
Thr (TGT)		+	
Trp (CCA)	+	+	
Tyr (GTA)	+	+	+
Val (CAC)	+		
Val (TAC)			+

Sulfolobus species (Tocchini-Valentini et al. 2005, Yoshinari et al. 2005). The catalytic subunit is conserved in size and sequence for the crenarchaea, but the structural subunit is more variable. *Hyperthermus butylicus* encodes the larger form (182 aa), whereas some crenarchaea, including *S. tokodaii* and *S. acidocaldarius*, encode a smaller protein (~92 aa) homologous to the C-terminal half of the larger protein (Brügger 2007).

Searches against the intein database were negative (Perler 2002). Moreover, an *A. pernix* ORF APE0745, carrying a predicted 468 aa intein encoding a homing endonuclease was ab-

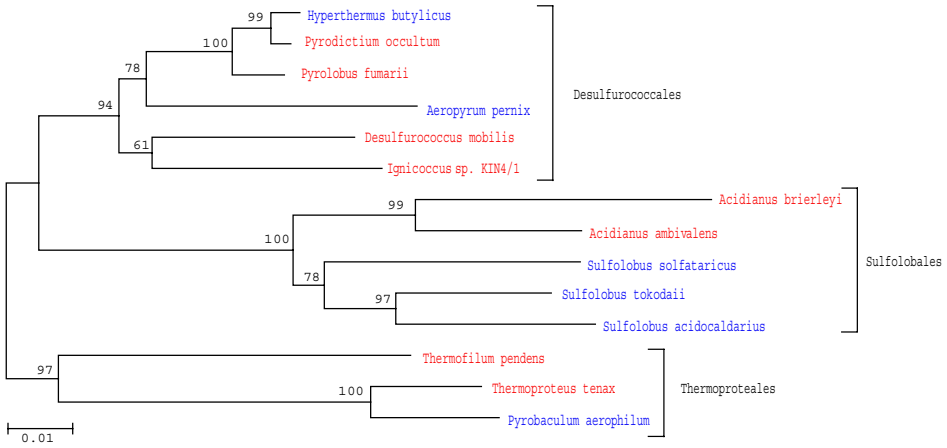

Figure 3. Phylogenetic tree based on 16S rRNA sequences generated with MEGA3 (Kumar et al. 2004) showing representatives of the three major crenarchaeal orders, Desulfurococcales (families *Desulfurococcaceae* and *Pyrodictiaceae*), Thermoproteales and Sulfolobales, for which completely sequenced genomes are available (blue) or for which genome sequences are incomplete and/or unpublished (red).

sent from the *H. butylicus* homologs (Hbut_0418, Hbut_0633 and Hbut_0856). We infer that *H. butylicus* lacks inteins as has been predicted for some other crenarchaea (e.g., Chen et al. 2005).

Phylogenetic status

The phylogenetic tree presented in Figure 3 was derived using 16S rRNA sequences. It illustrates the three major orders of the Crenarchaeota Desulfurococcales, Sulfolobales and Thermoproteales, for which several organisms have been purified and characterized (Stetter 2006). Included in the tree are those organisms (in blue) for which full genomic sequences have been published and those (in red) whose sequences are in progress or unpublished (Klenk 2006). Possibilities for comparative genomic studies of the crenarchaea are still limited because only six crenarchaeal genomes are available, three of which are from the *Sulfolobus* genus. As indicated in the Figure, the present genome, together with those which should soon be available, will give a strong basis for a comprehensive genomic comparison of the three best characterised orders of the kingdom Crenarchaeota.

Acknowledgments

Genome sequencing was supported by EU Cell Factory grant No. QLK3-CT-2000-00649, and further grants were received for an Archaea Centre supported by the Danish Natural Science Research Council. We thank Bettina Haberl (Epidauros Biotechnologie AG) and Hien Phan for help with DNA sequencing.

References

Aagaard, C., J. Dalgaard and R.A. Garrett. 1995. Inter-cellular mobility and homing of an archaeal rDNA intron confers selective advantage over intron- cells of *Sulfolobus acidocaldarius*. Proc. Natl. Acad. Sci. USA 92:12,285–12,289.

Baliga, N.S, R. Bonneau, M.T. Facciotti et al. 2004. Genome sequence of *Haloarcula marismortui*: a halophilic archaeon from the Dead Sea. Genome Res. 14:2221–2234.

Bateman, A., E. Birney, L. Cerruti et al. 2002. The Pfam protein families database. Nucl. Acids Res. 30:276–280.

Bendtsen, J.D., H. Nielsen, G. von Heijne and S. Brunak. 2004. Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. 340:783–795.

Benson, D.A., I. Karsch-Mizrachi, D.J. Lipman, J. Ostell and D.L. Wheeler. 2003. GenBank. Nucl. Acids Res. 31:23–27.

Boeckmann, B., A. Bairoch, R. Apweiler et al. 2003. The SWISS-PROT protein knowledge base and its supplement TrEMBL in 2003. Nucl. Acids Res. 31:365–370.

Brügger, K. 2007. The *Sulfolobus* database. Nucl. Acids Res. 35:D413–415.

Brügger, K., P. Redder, Q. She, F. Confalonieri, Y. Zivanovic and R.A. Garrett. 2002. Mobile elements in archaeal genomes. FEMS Microbiol. Lett. 206:131–141.

Brügger, K., P. Redder and M. Skovgaard. 2003. "MUTAGEN: Multi User Tool for Annotating Genomes" Bioinformatics 19: 2480–2481.

Brügger, K., X. Peng and R.A. Garrett. 2006. *Sulfolobus* genomes: mechanisms of rearrangement and change. *In* Archaea. Evolution, physiology and molecular biology. Eds. R.A. Garrett and H.-P. Klenk. Blackwell Publishing, Oxford, pp 95–104.

Burggraf, S., N. Larsen, C.R. Woese and K.O. Stetter. 1993. An intron within the 16S ribosomal RNA gene of the archaeon *Pyrobaculum aerophilum*. Proc. Natl. Acad. Sci. USA 90:2547–2550.

Cambillau, C. and J.-M.Claverie. 2000. Structural and genomic correlates of hyperthermostability. J. Biol. Chem. 43:32,383–32,386.

Chen, L., K. Brügger, M. Skovgaard et al. 2005. The genome of *Sulfolobus acidocaldarius*, a model organism of the *Crenarchaeota.* J. Bacteriol. 187:4992–4999.

Cobucci-Ponzano, B., A. Trincone, A. Giordano, M. Rossi and M. Moracci. 2003. Identification of an archaeal α-L-fucosidase encoded by an interrupted gene. Production of a functional enzyme by mutations mimicking programmed –1 frameshifting. J. Biol. Chem. 278:14,622–14,631.

Dirmeier, R., M. Keller, G. Frey, H. Huber and K.O. Stetter. 1998. Purification and properties of an extremely thermostable membrane-bound sulfur-reducing complex from the hyperthermophilic *Pyrodictium abyssi*. Eur. J. Biochem. 252:486–491.

Ewing, B. and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8: 186–194.

Ewing, B., L. Hillier, M. Wendl and P. Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8:175–185.

Fitz-Gibbon, S.T., H. Ladner, U-J. Kim, M.I. Simon and J.H. Miller. 2002. Genome sequence of the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum.* Proc. Natl. Acad. Sci. USA 99: 984–989.

Gogos, A. and N.D. Clarke. 1999. Characterization of an 8-oxoguanine DNA glycosylase from *Methanococcus jannaschii*. J. Biol. Chem. 274:3047–3045.

Gordon, D., C. Abajian and P. Green. 1998. Consed: a graphical tool for genome finishing. Genome Res. 8:195–202.

Greve, B., S. Jensen, K. Brügger, W. Zillig and R.A. Garrett. 2004. Genomic comparison of archaeal conjugative plasmids from *Sulfolobus*. Archaea 1:231–239.

Haft, D.H., J. Selengut, E.F. Mongodin and K.E. Nelson. 2005. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PloS Comput. Biol. 1:474–483.

Jansen, R., J.D. Embden, W. Gaastra and L.M. Schouls. 2002. Identification of genes that are associated with DNA repeats in prokaryotes. Mol. Microbiol. 43:1565–1575.

Kawarabayasi, Y., Y. Hino, H. Horikawa et al. 1999. Complete genome sequence of an aerobic hyperthermophilic crenarchaeon *Aeropyrum pernix* K1. DNA Res. 6:83–101.

Klenk, H.-P. 2006. Features of the genomes. *In* Archaea. Evolution, physiology and molecular biology. Eds. R.A. Garrett and H.-P. Klenk. Blackwell Publishing, Oxford, pp 75–94.

Kletzin, A. 2006. Metabolism of inorganic sulfur compounds in archaea. *In* Archaea. Evolution, physiology and molecular biology. Eds. R.A. Garrett and H.-P. Klenk. Blackwell Publishing, Oxford, pp 261–274.

Krogh, A., B. Larsson, G. von Heijne and E. Sonnhammer. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305: 567–580.

Kuchino, Y., M. Ihara, Y. Yabusaki and S. Nishimura. 1982. Initiator tRNAs from archaebacteria show common unique sequence characteristics. Nature 298:684–685.

Kumar, S., K. Tamura and M. Nei. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform. 5:150–163.

Larsen, T. and A. Krogh. 2003. EasyGene—a prokaryotic gene finder that ranks ORFs by statistical significance. BMC Bioinformatics 4:21.

Laska, S., F. Lottspeich and A. Kletzin. 2003. Membrane-bound hydrogenase and sulfur reductase of the hyperthermophilic and acidophilic archaeon *Acidianus ambivalens*. Microbiology 149: 2357–2371.

Lillestøl, R.K., P. Redder, R.A. Garrett and K. Brügger. 2006. A putative viral defence mechanism in archaeal cells. Archaea 2:59–72.

Lin, F.-H. and D.R. Forsdyke. 2007. Prokaryotes that grow optimally in acid have purine-poor codons in long open reading frames. Extremophiles 11:9–18.

Londei, P. 2006. Translational mechanisms and protein sysnthesis. *In* Archaea. Evolution, Physiology and Molecular Biology. Eds. R.A. Garrett and H.-P. Klenk. Blackwell Publishing, Oxford, pp 294–313.

Lowe, T. and S.R. Eddy. 1997. tRNAscan–SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucl. Acids Res. 25:955–964.

Lundgren, M., A. Andersson, L. Chen, P. Nilsson and R. Bernander. 2004. Three replication origins in *Sulfolobus* species: Synchronous initiation of chromosome replication and asynchronous termination. Proc. Natl. Acad. Sci. USA 101:7046–7051.

Makarova, K.S., L. Aravind, N.V. Grishin, I.B. Rogozin and E.V. Koonin. 2002. A DNA repair system specific for thermophilic archaea and bacteria predicted by genome context analysis. Nucl. Acids Res. 30:482–496.

Makarova, K.S., N.V. Grishin, S.A. Shabalina, Y.I. Wolf and E.V. Koonin. 2006. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biology Direct 1:7.

Marck, C. and H. Grosjean. 2003. Identification of BHB splicing motifs in intron-containing tRNAs from 18 archaea: evolutionary implications. RNA 9:1516–1531.

Marsh, V.L. and S.D. Bell. 2006. DNA replication and the cell cycle. *In* Archaea. Evolution, physiology and molecular biology. Eds. R.A. Garrett and H.-P. Klenk. Blackwell Publishing, Oxford, pp 217–231.

Matsunaga, F., P. Forterre, Y. Ishino and H. Myllykallio. 2001. Interactions of archaeal Cdc6/Orc1 and minichromosome maintenance proteins with the replication origin. Proc. Natl. Acad. Sci. USA 98:11,152–11,157.

Mojica, F.J., C. Diez-Villasenor, J. Garcia-Martinez and E. Soria. 2005. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J. Mol. Evol. 60: 174–182.

Ng S.Y., B. Chaban and K.F. Jarrell. 2006. Archaeal flagella, bacterial flagella and type IV pili: a comparison of genes and posttranslational modifications. J. Mol. Microbiol. Biotechnol. 11:167–191.

Peng, X., K. Brügger, B. Shen, L. Chen, Q. She and R.A. Garrett. 2003. Genus-specific protein binding to the large clusters of DNA repeats (Short Regularly Spaced Repeats) present in *Sulfolobus* genomes. J. Bacteriol. 185:2410–2417.

Perler, F.B. 2002. InBase, the intein database. Nucl. Acids Res. 30:383–384.

Prangishvili, D. and R.A. Garrett. 2005. Viruses of hyperthermophilic crenarchaea. Trends Microbiol. 13:535–542.

Randau, L., R. Münch, M.J. Hohn, D. Jahn and D. Söll. 2005. *Nanoarchaeum equitans* creates functional tRNAs from separate genes for their 5′-and 3′-halves. Nature 433:537–541.

Redder, P., Q. She and R.A. Garrett. 2001. Non-autonomous mobile elements in the Crenarchaeon *Sulfolobus solfataricus*. J. Mol. Biol. 306:1–6.

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

Rozenski, J., P.F. Crain and J.A. McCloskey. 1999. The RNA Modification Database. Nucl. Acids Res. 27:196–197.

She, Q., R.K. Singh, F. Confalonieri et al. 2001. The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. Proc. Natl. Acad. Sci. USA 98:7835–7840.

She, Q., B. Shen and L. Chen. 2004. Archaeal integrases and mechanisms of gene capture. Biochem. Soc. Trans. 32:222–226.

Stetter, K.O. 2006. History of the discovery of the first hyperthermophiles. Extremophiles 10:357–362.

Tang, T.-H., J-P. Bachellerie, T. Rozhdestvensky, M-L. Bortolin, H. Huber, M. Drungowski, T. Elge, J. Brosius and A. Hüttenhofer. 2002. Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. Proc. Natl. Acad. Sci. USA 99:7536–7541.

Tang, T.-H., N. Polacek, M. Zywicki, H. Huber, K. Brügger, R. Garrett, J.P. Bachellerie and A. Hüttenhofer. 2005. Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus.* Mol. Microbiol. 55:469–481.

Tatusov, R.L., D.A. Natale, I.V. Garkavtsev et al. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucl. Acids. Res. 29:22–28.

Tocchini-Valentini, G.D., P. Fruscoloni and G.P. Tocchini-Valentini. 2005. Structure, function and evolution of the tRNA endonucleases of Archaea: an example of subfunctionalization. Proc. Nat. Acad. Sci. USA 102:8933–8938.

Torarinsson, E., H.-P. Klenk and R.A. Garrett. 2005. Divergent transcriptional and translational signals in Archaea. Environ. Microbiol. 7:47–54.

Toth, J., A.A. Ismaiel and J.S. Chen. 1999. The *ald* gene, encoding a coenzyme A-acylating aldehyde dehydrogenase, distinguishes *Clostridium beijerinckii* and two other solvent-producing clostridia from *Clostridium acetobutylicum*. Appl. Environ. Microbiol. 65:4973–4980.

Watanabe, Y., S. Yokobori, T. Inaba, A. Yamagishi, T. Oshima, Y. Kawarabayashi, H. Kikuchi and K. Kita. 2002. Introns in protein coding genes in archaea. FEBS Lett. 510:27–30.

White, M.F. 2006. DNA repair. *In* Archaea. Evolution, physiology and molecular biology. Eds. R.A. Garrett and H.-P. Klenk. Blackwell Publishing, Oxford, pp 171–184.

Yang, W. 2005. Portraits of a Y-family DNA polymerase. FEBS Lett. 579:868–872.

Yoshinari, S., S. Fujita, R. Masui, S. Kuramitsu, S. Yokobori, K. Kita and Y. Watanabe. 2005. Functional reconstitution of a crenarchaeal splicing endonuclease in vitro. Biochem. Biophys. Res. Comm. 334:1254–1259.

Yoshinari, S., T. Itoh, S.J. Hallam, E.F. Delong, S. Yokobori, A. Yamagishi, T. Oshima, K. Kita and Y. Watanabe. 2006. Archaeal pre-mRNA splicing: A connection to hetero-oligomeric splicing endonuclease. Biochem. Biophys. Res. Comm. 346:1024–1032.

Zhang, R. and C. Zhang. 2005. Identification and replication origins in archaeal genomes based on the Z-curve method. Archaea 1: 335–346.

Zillig, W., I. Holz, D. Janekovic et al. 1990. *Hyperthermus butylicus*, a hyperthermophilic sulfur-reducing archaebacterium that ferments peptides. J. Bacteriol. 172:3959–3965.

Zillig W., I. Holz, and S. Wunderl. 1991. *Hyperthermus butylicus* gen. nov., sp. nov., a hyperthermophilic, anaerobic, peptide-fermenting, facultatively H_2S-generating archaebacterium. Int. J. Syst. Bacteriol. 41:169–170.

Zillig, W., H.P. Arnold, I. Holz et al. 1998. Genetic elements in the extremely thermophilic archaeon *Sulfolobus*. Extremophiles 2: 131–140.