

## Complete Genome Sequence of an Aerobic Hyper-thermophilic Crenarchaeon, *Aeropyrum pernix* K1

Yutaka KAWARABAYASI,<sup>1,2,\*</sup> Yumi HINO,<sup>1</sup> Hiroshi HORIKAWA,<sup>1</sup> Syuji YAMAZAKI,<sup>1</sup> Yuji HAIKAWA,<sup>1</sup> Koji JIN-NO,<sup>1</sup> Mikio TAKAHASHI,<sup>1</sup> Mitsuo SEKINE,<sup>1</sup> Sin-ichi BABA,<sup>1</sup> Akiho ANKAI,<sup>1</sup> Hiroki KOSUGI,<sup>1</sup> Akira HOSOYAMA,<sup>1</sup> Shigehiro FUKUI,<sup>1</sup> Yoshimi NAGAI,<sup>1</sup> Keiko NISHIJIMA,<sup>1</sup> Hidekazu NAKAZAWA,<sup>1</sup> Minako TAKAMIYA,<sup>1</sup> Sayaka MASUDA,<sup>1</sup> Tomomichi FUNAHASHI,<sup>1</sup> Toshihiro TANAKA,<sup>1</sup> Yutaka KUDOH,<sup>1</sup> Jun YAMAZAKI,<sup>1</sup> Norihiro KUSHIDA,<sup>1</sup> Akio OGUCHI,<sup>1</sup> Ken-ichi AOKI,<sup>1</sup> Kenji KUBOTA,<sup>1</sup> Yoshinobu NAKAMURA,<sup>1</sup> Norimichi NOMURA,<sup>3</sup> Yoshihiko SAKO,<sup>3</sup> and Hisasi KIKUCHI<sup>1</sup>

National Institute of Technology and Evaluation, 2-49-10 Nishihara, Shibuya, Tokyo 151-0066, Japan,<sup>1</sup> National Institute of Bioscience and Human-Technology, 1-1 Higashi, Tsukuba, Ibaraki 305-0046, Japan,<sup>2</sup> and Kyoto University, Sakyo-ku, Kyoto 606-8502, Japan<sup>3</sup>

(Received 9 March 1999)

### Abstract

The complete sequence of the genome of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1, which optimally grows at 95°C, has been determined by the whole genome shotgun method with some modifications. The entire length of the genome was 1,669,695 bp. The authenticity of the entire sequence was supported by restriction analysis of long PCR products, which were directly amplified from the genomic DNA. As the potential protein-coding regions, a total of 2,694 open reading frames (ORFs) were assigned. By similarity search against public databases, 633 (23.5%) of the ORFs were related to genes with putative function and 523 (19.4%) to the sequences registered but with unknown function. All the genes in the TCA cycle except for that of alpha-ketoglutarate dehydrogenase were included, and instead of the alpha-ketoglutarate dehydrogenase gene, the genes coding for the two subunits of 2-oxoacid:ferredoxin oxidoreductase were identified. The remaining 1,538 ORFs (57.1%) did not show any significant similarity to the sequences in the databases. Sequence comparison among the assigned ORFs suggested that a considerable number of ORFs were generated by sequence duplication. The RNA genes identified were a single 16S–23S rRNA operon, two 5S rRNA genes and 47 tRNA genes including 14 genes with intron structures. All the assigned ORFs and RNA coding regions occupied 89.12% of the whole genome. The data presented in this paper are available on the internet homepage (<http://www.mild.nite.go.jp>).

**Key words:** aerobic hyper-thermophile; crenarchaeota; *Aeropyrum pernix* K1; genome sequencing; whole genome shotgun method

### 1. Introduction

The complete genome sequences of four kinds of thermophilic and anaerobic archaea have been reported,<sup>1–4</sup> including the one sequenced by our group (*Pyrococcus horikoshii* OT3).<sup>1</sup> Among these archaea, *Pyrococcus horikoshii* OT3 is a hyper-thermophilic strain with an optimal growth temperature of 98°C, whereas the other three strains grow in the range of 60 to 85°C. To accumulate the information on the genetic system of or-

ganisms living at hyper-thermophilic conditions, we now analysed the genome of another archaeon, *Aeropyrum pernix* K1. This strain has been isolated in 1993 from coastal solfataric thermal vent in Kodakara-jima Island in Kyusyu, Japan.<sup>5</sup> Like *P. horikoshii* OT3, it grows in the range of 90 to 98°C with an optimal temperature of 95°C. However, compared with *P. horikoshii* OT3 that belongs to an anaerobic euryarchaeota, *A. pernix* K1 is an aerobic strain classified as crenarchaeota. Therefore, the sequence data on its genome should also provide information on the origin of the oxygenic respiratory system.

To sequence the entire genome, we first constructed the restriction maps of the genome, using three kinds of 8-base cutters, *Pac* I, *Sgf* I and *Swa* I. Then the shotgun libraries with short and long inserts were constructed from both the entire genomic DNA and restriction fragments

Communicated by Mituru Takanami

\* To whom correspondence should be addressed. Tel. +81-3-3481-8951, Fax. +81-3-3481-8424, E-mail: kiyutaka@kazusa.or.jp

† The entire genome sequence has been deposited in DDBJ/Genbank/EMBL databases under the accession numbers AP000058-AP000064.

generated by *Sgf*I and *Swa*I. The single-run sequence data of the shotgun clones were accumulated, and the entire genomic sequence was directly deduced by assembling the sequence data with the software PhredPhrap,<sup>6,7</sup> followed by gap-filling by walking with long-insert shotgun clones.

## 2. Materials and Methods

### 2.1. Bacterial strains and genomic DNA

The *A. pernix* K1 strain deposited in the Japan Collection of Microorganisms (JCM number 9820) was used for genome sequencing in this study. The cells were inoculated into 100 ml of the JXT medium<sup>5</sup> which was prepared in 500-ml Erlenmeyer flasks, and cultured at 90°C with vigorous shaking. The genomic DNA was isolated principally based on the method of Sako et al.<sup>5</sup> *Escherichia coli* DH10B was used for the preparation of plasmid clones.

### 2.2. Construction of shotgun clones

The *A. pernix* K1 genomic DNA was sonicated to three different extents (5, 10 and 20 sec at the L position of Biorupter; Cosmo Bio, Tokyo, Japan). Followed by size-fractionation with agarose gel electrophoresis, the fractions 0.8 to 1.2 kb and 2.0 to 2.5 kb were independently cloned into the *Hinc* II site of pUC118 (short and long fragment-shotgun libraries).

For the construction of the shotgun libraries from restriction fragments, the genomic DNA was prepared in the agarose-plug as described by Smith et al.,<sup>8</sup> and after complete digestion of DNA in the plugs with *Sgf*I or *Swa*I, the digests were resolved by 0.4% agarose gel electrophoresis. The band regions were cut out, and agarose gel was dissolved by incubation overnight at 50°C with Agarase (FMC, Rockland, ME, USA). The resulting solution containing DNA fragments was sonicated, and the short fragment-shotgun libraries were prepared as above, except for *Sgf*I fragment A (1015 kb) from which both the short and long fragment-shotgun libraries were constructed.

### 2.3. DNA sequencing

Plasmid DNA in each shotgun library was prepared by the Autogen 740 automatic DNA preparation system (Autogen, Framington, MA, USA). The sequencing reaction was performed using two kinds of cycle sequencing kits, dye-primer cycle sequencing kit and dye-terminator cycle sequencing kit, and ABI-DNA sequencers (373XL and 377XL; Perkin-Elmer ABI, Foster City, CA, USA). For the sequence determination at the gaps, the clones bridging two contigs were screened from the long fragment-shotgun libraries, and sequenced by the dye-terminator method.

### 2.4. Data assembly and computational analysis

The accumulated files of single-run sequence data were first treated using the software Phred<sup>6</sup> to eliminate any contaminated sequences derived from *E. coli* or vector DNAs, and the treated data were assembled into contigs using the software Phrap.<sup>7</sup> The assembled sequences were split into 30-kbp segments, and the sequence in each segment was re-assembled and edited by Sequencher (Gene Codes, Ann Arbor, MI, USA).

The criteria used for the assignment of potential coding regions on the genomic sequence are as follows. All the ORFs of larger than 100 sense codons starting with ATG or GTG were first assigned, and then ORFs of 50 to 99 sense codons were taken, if the sequences showed any similarity to the coding sequences in the databases or to the protein motif sequences. Similarity search of assigned ORFs was performed by the Smith-Waterman algorithm.<sup>9</sup> The databases used for similarity search were Genbank release 109, EMBL release 56.0, Swiss-Prot release 36.0, PIR release 57.0 and Owl release 31.0.

## 3. Results and Discussion

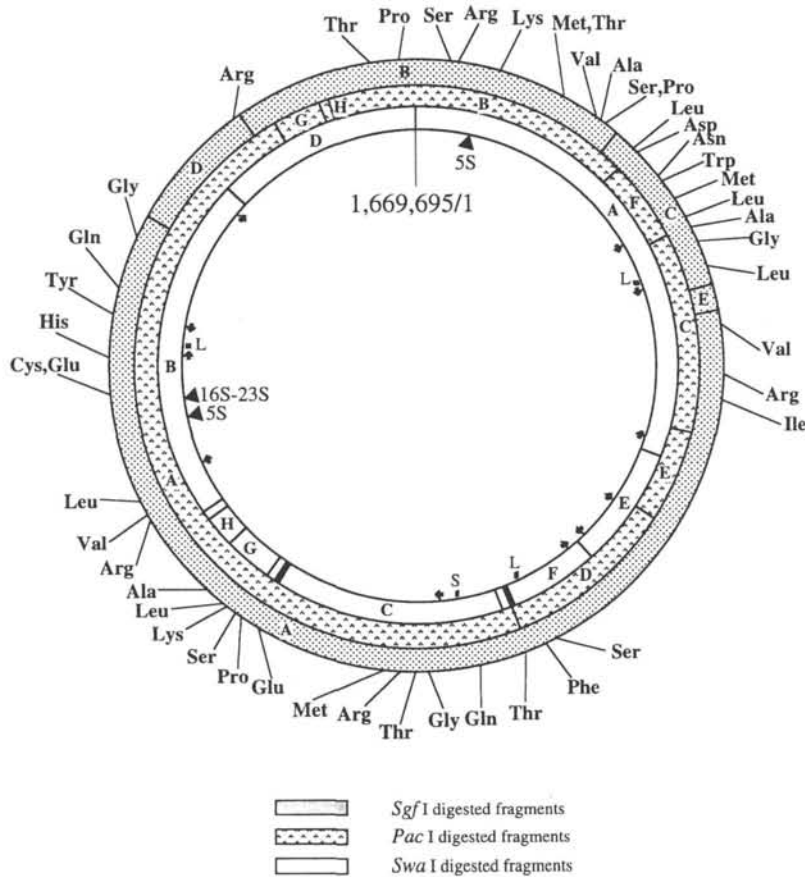
### 3.1. Construction of physical map

The genomic DNA of *A. pernix* K1 was digested with three kinds of 8-base cutters, *Sgf*I, *Pac*I and *Swa*I, which generated 5, 8 and 13 fragments, respectively. The resulting fragments were resolved and purified by the CHEF Mapper gel electrophoresis system (BioRad, Richmond, CA, USA). Then the fragments were ordered by hybridization analysis among the fragments and their double-digestion products.

Figure 1 shows the physical maps which were finally confirmed by the determination of the entire genomic sequence. Three of the *Swa*I sites which were present in the close vicinity of other *Swa*I sites were assigned on the basis of the deduced sequence. Fragments A to E generated by *Sgf*I and fragments F to H generated by *Swa*I fragments were used for the construction of restriction fragment-specific shotgun libraries.

### 3.2. Determination of entire genome sequence

For deduction of the entire genome sequence, two kinds of shotgun libraries, one with approximately 1-kb inserts and the other with approximately 2.2-kb inserts, were prepared from both the genomic DNA and restriction fragments as mentioned above. The clones from the short fragment-shotgun libraries were single-run sequenced from one end and those from the long fragment-shotgun libraries from both the ends. A total of 29,000 single-run sequence data, corresponding to approximately 8 times the entire genomic length, were accumulated and used for assembling by PhredPhrap. The contigs primarily assembled by PhredPhrap constituted 14 islands. For the gap-filling, the clones carrying



**Figure 1.** The physical map of the *A. pernix* K1 genome and location of RNA genes and characteristic repeating units. The restriction sites for *Sgf* I, *Pac* I and *Swa* I are indicated in the circular maps of the genome. The nucleotide positions are numbered in the clockwise direction from the junction of the *Swa* I A and D fragments (the position is given as 1,669,695/1). The positions and species of the tRNA genes are indicated by bars with the names of chargeable amino acids. The position of the 5S and 16S–23S rRNA genes are shown by triangles labeled with 5S and 16–23S, and those of type A and type B repeating units by boxes labeled L and S, respectively. The arrows along the inner circle indicate the position and direction of the type C repeating units.

the end sequences of contigs were screened from the long fragment-shotgun libraries and sequenced. All the gaps were closed by sequencing two clones per gap. The gap lengths finally confirmed were 59 to 553 bp. The criterion taken for the sequence confirmation was coincidence of the sequences deduced from different shotgun clones.

To confirm the authenticity of the genomic sequence constructed, every 15-kb region of the genome was directly amplified from the genomic DNA by long PCR, and the occurrence of restriction sites were compared with those deduced from the sequence data.

In the course of editing, it was noted that different nucleotide signals appear at the same position and that the results were reproducible. Such positions were identified at at least 35 different positions along the entire genome: the mixture of G and A signals at 15 positions, that of C and T signals at 9 positions and that of other base signals at 11 positions. The occurrence of different nucleotides at the same position could be resolved by sequencing of the PCR products directly amplified from the genomic DNA. We assume that this is due to the presence of multiple

copies of the genome in a single cell. In the sequence registered in the databases, the nucleotides giving the major signals are indicated.

The total length of the genome finally confirmed was 1,669,695 bp. The nucleotide position was numbered from the one end of the *Swa* I restriction site located on the *Sgf* I B and *Pac* I B fragments, as shown in the physical map in Fig. 1. The distribution of ACGT along the strand of the entire genome was 21.6% A, 28.4% C, 28.0% G and 22.1% T, resulting in a G + C content of 56.3%. This is the highest G + C content among the genomes of four archaea of which the entire genomic sequence were reported. Through the processes of physical mapping and genomic sequencing, there was no evidence of extra-chromosomal units.

### 3.3. Assignment and similarity search of potential protein-coding regions

The assignment of potential protein-coding regions was done as in the previous paper.<sup>1</sup> The ORFs which consist

**Table 1.** Codon-usages of the predicted proteins coded for by the 2,618 APE-class ORFs.

UUU (Phe)	6.01	3823	UCU (Ser)	9.07	5767	UAU (Tyr)	12.65	8042	UGU (Cys)	3.20	2037
UUC (Phe)	21.35	13573	UCC (Ser)	15.55	9886	UAC (Tyr)	20.80	13225	UGC (Cys)	6.07	3859
UUA (Leu)	4.74	3015	UCA (Ser)	7.00	4448	UAA (Stop)	0.84	531	UGA (Stop)	0.90	573
UUG (Leu)	7.16	4554	UCG (Ser)	10.82	6880	UAG (Stop)	2.38	1514	UGG (Trp)	13.09	8323
CUU (Leu)	16.85	10712	CCU (Pro)	17.96	11420	CAU (His)	6.11	3883	CGU (Arg)	3.75	2385
CUC (Leu)	38.48	24467	CCC (Pro)	24.18	15373	CAC (His)	13.06	8306	CGC (Arg)	6.75	4291
CUA (Leu)	18.00	11443	CCA (Pro)	9.08	5776	CAA (Gln)	3.04	1934	CGA (Arg)	2.63	1671
CUG (Leu)	28.12	17879	CCG (Pro)	13.09	8323	CAG (Gln)	15.92	10123	CGG (Arg)	6.99	4446
AUU (Ile)	7.92	5036	ACU (Thr)	8.94	5682	AAU (Asn)	4.49	2852	AGU (Ser)	6.72	4276
AUC (Ile)	12.44	7911	ACC (Thr)	16.16	10278	AAC (Asn)	15.81	10052	AGC (Ser)	25.82	16418
AUA (Ile)	31.36	19942	ACA (Thr)	10.57	6720	AAA (Lys)	7.71	4900	AGA (Arg)	11.16	7099
AUG (Met)	19.62	12475	ACG (Thr)	11.03	7012	AAG (Lys)	27.40	17426	AGG (Arg)	45.48	28922
GUU (Val)	20.98	13343	GCU (Ala)	24.76	15743	GAU (Asp)	12.62	8022	GGU (Gly)	14.87	9457
GUC (Val)	22.62	14383	GCC (Ala)	37.01	23533	GAC (Asp)	26.00	16536	GGC (Gly)	32.66	20766
GUA (Val)	14.62	9296	GCA (Ala)	13.51	8592	GAA (Glu)	11.53	7331	GGA (Gly)	11.98	7616
GUG (Val)	28.93	18395	GCG (Ala)	19.67	12510	GAG (Glu)	54.34	34556	GGG (Gly)	25.66	16318

Numerals in the second column indicate the frequency of occurrence per thousand, and those in the third column the sum of codons occurred.

of longer than 100 sense codons starting with ATG or GTG were first searched, and are represented by a three-letter code (APE) plus a four-digit number indicating the ORF position. There was a total of 2,618 of this class of ORF including the overlapped ORFs in the same or either strand. In the next, shorter ORFs consisting of 50 to 99 codons were extracted from the regions where no ORF of longer than 100 codons was assigned, and the ORFs which possess significant similarity to either the registered sequences or protein motifs were taken as the probable protein-coding regions. Those were named with a four-letter code (APES) plus a two-digit number indicating the ORF position.

The ORFs thus assigned along the entire genomic sequence were totalled to 2,694, (APE-class ORFs = 2,618, APES-class ORFs = 76). An average size of the ORFs was 237 amino acid residues, and the longest one was consisted of 1,933 residues (APE0620). The assigned 2,694 ORFs occupy 88.83% of the genome.

For the anticipation of gene function, the products with a Zscore of  $\geq 25$  or with over 30% amino acid sequence identity along the entire coding region were taken into account for the APE-class ORFs, and the ZScore of  $\geq 10$  was used only for the APES-class ORFs. Among the ORFs assigned, the number of ORFs related to the genes with functions were 633 (23.5%), ORFs showed significant similarity to the registered sequences with unknown functions were 523 (19.4%) and those just containing some motifs were 202 (7.5%). The remaining 1,336 ORFs showed no significant similarity to the sequences in public databases.

The ORFs showing similarities to the registered genes with a known function and these containing protein-motifs are listed in the last two figures (Tables 5 and

6). The codon-usages of the hypothetical proteins coded by the total APE-class ORFs are summarized in Table 1. The codons with G or C at the third position appeared to be more frequently used, probably reflecting the relatively high G + C content of this strain.

### 3.4. RNA coding genes

The entire genomic sequence was subjected to a similarity search against the rRNA sequences registered in the databases, and a single 16S–23S rRNA cluster and two 5S rRNA genes were assigned, as shown in Fig. 1. Such an organization is similar to that of *A. fulgidus*<sup>4</sup> and *P. horikoshii* OT3.<sup>1</sup> It has been reported that both the 16S and 23S rRNA genes of *A. pernix* K1 contain introns<sup>10</sup> and that the intron in the 16S rRNA gene encodes a homing DNA endonuclease.<sup>10</sup> We noted that an ORF similar to the DNA endonuclease gene is present in one intron of the 23S rRNA gene (APE1920 in Table 5).

By searching with tRNAscan,<sup>11</sup> 47 tRNA genes were identified, 14 of which were found to contain 18- to 121-bp introns. The species of tRNA genes identified are shown in Table 2, and those containing introns and the organization of the introns within each gene are summarized in Table 3. In all of the intron-containing tRNA genes except for the Thr- and Trp-tRNA genes, the introns were located one base from the anticodon region. Forty-one of them were discretely mapped, while the remaining six tRNA genes were mapped as clusters of two tRNA genes. As noted in other microorganisms,<sup>1,12</sup> no tRNA genes containing A at the first position of the anticodon were present. As in all the archaea genomes sequenced, the Met-tRNA gene was present in triplicate. However, the sequences had no similarity each other, suggesting that the different Met-tRNA species may be used for the

Table 2. Summary of tRNA gene assignment.

UUU (Phe)		UCU (Ser)		UAU (Tyr)		UGU (Cys)	
UUC (Phe)	○	UCC (Ser)	○	UAC (Tyr)	⊙	UGC (Cys)	⊙
UUA (Leu)	○	UCA (Ser)	○	UAA (Stop)		UGA (Stop)	
UUG (Leu)	○	UCG (Ser)	⊙	UAG (Stop)		UGG (Trp)	⊙
CUU (Leu)		CCU (Pro)		CAU (His)		CGU (Arg)	
CUC (Leu)	○	CCC (Pro)	⊙	CAC (His)	○	CGC (Arg)	○
CUA (Leu)	○	CCA (Pro)	○	CAA (Gln)	○	CGA (Arg)	○
CUG (Leu)	○	CCG (Pro)	⊙	CAG (Gln)	○	CGG (Arg)	○
AUU (Ile)		ACU (Thr)		AAU (Asn)		AGU (Ser)	
AUC (Ile)	○	ACC (Thr)	○	AAC (Asn)	○	AGC (Ser)	○
AUA (Ile)		ACA (Thr)	⊙**	AAA (Lys)	⊙	AGA (Arg)	⊙
AUG (Met)	⊙*	ACG (Thr)	⊙	AAG (Lys)	⊙	AGG (Arg)	○
GUU (Val)		GCU (Ala)		GAU (Asp)		GGU (Gly)	
GUC (Val)	○	GCC (Ala)	○	GAC (Asp)	⊙	GGC (Gly)	○
GUA (Val)	○	GCA (Ala)	○	GAA (Glu)	○	GGA (Gly)	○
GUG (Val)	○	GCG (Ala)	○	GAG (Glu)	○	GGG (Gly)	○

The tRNA genes identified are indicated by circles and those containing introns by double-circles in the corresponding triple boxes.

\*: Two of the three Met-tRNA genes possess the introns.

\*\* : One of the two Thr-tRNA genes possesses the intron.

Table 3. The size and position of introns in the intron-containing tRNA genes.

tRNA species	anticodon	gene size (nt)	intron position*		intron size (nt)	mature tRNA (nt)
			start	stop		
Arg-tRNA	TCT	122	40	83	44	78
Asp-tRNA	GTC	199	41	161	121	78
Cys-tRNA	GCA	96	39	56	18	78
Lys-tRNA	CTT	113	39	72	34	79
Lys-tRNA	TTT	113	39	72	34	79
Met-tRNA	CAT	115	39	76	38	77
Met-tRNA	CAT	125	39	86	48	77
Pro-tRNA	GGG	115	40	76	37	78
Pro-tRNA	CGG	122	40	83	44	78
Ser-tRNA	CGA	125	39	74	36	79
Thr-tRNA	CGT	124	37	85	49	75
Thr-tRNA	TGT	95	22	40	19	74
Tyr-tRNA	GTA	114	37	75	39	75
Trp-tRNA	CCA	114	32	68	37	77

\*: Nucleotide positions from the 5' end of tRNA genes.

translation initiation of different classes of genes.

The genes for all the corresponding tRNA synthetases have been identified in the assigned ORFs, except for those of Gln-tRNA and Asn-tRNA. Instead, the three subunits of Glu-tRNA amidotransferases, which are necessary for Gln- and Asn-tRNA synthesis, were identified.

### 3.5. Other features noted

#### 3.5.1. ORF containing intein

The presence of inteins, defined as the self-splicing portions of a polypeptide sequence,<sup>13,14</sup> has been reported

in 36 archaea genes.<sup>1-3,15-19</sup> To investigate the possible occurrence of ORFs containing inteins in the genome of *A. pernix* K1, all the ORF sequences were subjected to a similarity search against the intein-specific motif sequences described by Pietrokovski<sup>20</sup> and Perler,<sup>21</sup> and also the amino acid sequences in the reported inteins using the Smith-Waterman algorithm. According to the results, one ORF (APE0745) was found to contain a putative intein element. The ORF length was 726 amino acids, of which the intein element occupied 468 amino acids. However, this ORF belonged to the ORF class coding for hypothetical proteins.

**Table 4.** Homologous ORFs identified by sequence comparison among the 2618 APE-class ORFs.

number of homologous ORFs*	number of groups	total ORFs
10	2	20
9	2	18
8	3	24
7	6	42
6	6	36
5	7	35
4	7	28
3	25	75
2	127	254
	total	532

\*: ORFs with ZScore higher than 8 in SW search and amino acid identity higher than 30% in 70% of the entire region were taken.

### 3.5.2. Duplication of ORFs

By sequence comparison among the ORFs assigned to the genome, 532 ORFs could be grouped into 185, the ORFs in each group coding for proteins with considerable similarity (Table 4). The result can be interpreted that the ORFs in each group were generated by duplication of an ancestral sequence. Such homologous ORFs were present in the genome as either tandem repeats of a single ORF or the repeats of a single or a cluster of ORFs at different locations.

### 3.5.3. Unique repeating sequences

Three types of repeating units were found on the *A. pernix* K1 genome by similarity search, while no IS- or Tn-like elements were identified. The repeating unit composed of a well-conserved 24 bp sequence (GAATCTTC-GAGATAGAATTGCAAG) followed by a variable 37–52 bp sequence and 366- or 365-bp repeating units was classified as a type A repeating unit. This type of the repeating unit was present at three different positions, as indicated in Fig. 1 (positions 325,586–325,975, 716,882–718,997, and 1,277,298–1,278,859). Type B repeating unit was composed of 42 times-repeating well-conserved 23-bp sequence (ATATCCCTAAAGGGAATAGAAAG) followed by a variable 38–49 bp sequence. This was present at a single position (786,659–789,355). The type C repeating unit was merely dispersion of approximately 300-bp repeats along the entire genome (see Fig. 1). The biological significance of these repeats is not known.

### 3.5.4. Genes relating to respiration

Since *A. pernix* K1 is strictly aerobic, the genes involved in the TCA cycle were first searched, and all of the genes in this cycle except for alpha-ketoglutarate dehydrogenase were found. Instead of the genes for alpha-ketoglutarate dehydrogenase, the genes coding for the two subunits of 2-oxoacid:ferredoxin oxidoreductase,

which possesses the same function in archaea as alpha-ketoglutarate dehydrogenase,<sup>22</sup> was identified on the genome. This suggests that *A. pernix* K1 had already acquired a respiratory system similar to that seen in mitochondria of higher eukaryotes.

As other genes relating to the respiratory chain, those for NADH dehydrogenase, NADH-quinone oxidoreductase, succinate dehydrogenase, iron-sulfur protein, flavo-protein, cytochrome *b* and cytochrome *c* oxidase were identified, and one ORF was found to have weak homology with the Rieske protein. However, none of the genes coding for quinol-cytochrome *c* oxidoreductase, cytochrome *c* reductase, cytochrome *c*<sub>1</sub>, cytochrome *bc*<sub>1</sub>, cytochrome *b*<sub>6f</sub>, cytochrome *c* and quinol oxidase, were identified by similarity search.

Like all other aerobic organisms the gene for superoxide dismutase was also present in this organism.

Analysis of gene organization as well as gene structure of the *A. pernix* K1 genome in comparison with those of the *P. horikoshii* OT3 genome is under investigation to find any clue for understanding the thermostability of the gene products.

**Acknowledgments:** We thank Dr. S. Tabata and H. Kotani and their staffs at Kazusa DNA Research Institute for their valuable suggestions and help. Thanks are also due to Dr. D. Söll for his kind help with the assignment of tRNA genes and to Dr. P. Green for donation of the assembling software PhredPhrap. This work was supported by the Ministry of International Trade and Industry.

## References

1. Kawarabayasi, Y., Sawada, M., Horikawa, H. et al. 1998, Complete Sequence and Gene Organization of the Genome of a Hyper-thermophilic Archaeobacterium, *Pyrococcus horikoshii* OT3, *DNA Res.*, **5**, 55–76 & 147–155.
2. Bult, C. J., White, O., Olsen, G. J. et al. 1996, Complete Genome Sequence of the Methanogenic Archaeon, *Methanococcus jannaschii*, *Science*, **273**, 1058–1073.
3. Smith, D. R., Doucette-Stamm, L. A., Deloughery, C. et al. 1997, Complete Genome Sequence of *Methanobacterium thermoautotrophicum* ΔH: Functional Analysis and Comparative Genomics, *J. Bacteriol.*, **179**, 7135–7155.
4. Klenk, H. P., Clayton, R. A., Tomb, J. F. et al. 1997, The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*, *Nature*, **390**, 364–370.
5. Sako, Y., Nomura, N., Uchida, A. et al. 1996, *Aeropyrum pernix* gen. nov., sp. nov., a Novel Aerobic Hyperthermophilic Archaeon Growing at Temperatures up to 100°C, *International Journal of Systematic Bacteriology*, **46**, 1070–1077.
6. Ewing, B., Hillier, L., Wendl, M. C., and Green, P. 1998, Base-calling of Automated Sequencer Traces Using phred. I. Accuracy Assessment, *Genome Res.*, **8**, 175–185.

7. Ewing, B. and Green, P. 1998, Base-calling of Automated Sequencer Traces Using phred. II. Error Probabilities, *Genome Res.*, **8**, 186–194.
8. Smith, C. L., Econome, J. G., Schutt, A., Klco, A., and Cantor, C. R. 1987, A Physical Map of the *Escherichia coli* K12 Genome, *Science*, **236**, 1448–1453.
9. Smith, T. F. and Waterman, M. S. 1981, Identification of Common Molecular Subsequences, *J. Mol. Biol.*, **147**, 195–197.
10. Nomura, N., Sako, Y., and Uchida, A. 1998, Molecular Characterization and Postsplicing Fate of Three Introns within the Single rRNA Operon of the Hyperthermophilic Archaeon *Aeropyrum pernix* K1, *J. Bacteriol.*, **180**, 3635–3643.
11. Lowe, T. M. and Eddy, S. R. 1997, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nuc. Acids Res.*, **25**, 955–964.
12. Nakamura, Y. and Tabata, S. 1997, Codon-Anticodon Assignment and Detection of Codon Usage Trends in Seven Microbial Genomes, *Microbial & Comparative Genomics*, **2**, 299–312.
13. Hirata, R., Ohsumi, Y., Nakano, A., Kawasaki, H., Suzuki, K., and Anraku, Y. 1990, Molecular Structure of a Gene, VMA1, Encoding the Catalytic Subunit of H<sup>+</sup>-Translocating Adenosine Triphosphatase from Vacuolar Membranes of *Saccharomyces cerevisiae*, *J. Biol. Chem.*, **265**, 6726–6733.
14. Kane, P. M., Yamashiro, C. T., Wolczyk, D. F., Neff, N., Goebel, M., and Stevens, T. H. 1990, Protein Splicing Converts the Yeast TFP1 Gene Product to the 69-kD Subunit of the Vacuolar H<sup>+</sup>-Adenosine Triphosphatase, *Science*, **250**, 651–657.
15. Riera, J., Robb, F. T., Weiss, R., and Fontecave, M. 1990, Ribonucleotide reductase in the archaeon *Pyrococcus furiosus*: a critical enzyme in the evolution of DNA genomes?, *Proc. Natl. Acad. Sci. USA*, **94**, 475–478.
16. Xu, M., Southworth, M. W., Mersha, F. B., Hornstra, L. J., and Peler, F. B. 1993, In vitro protein splicing of purified precursor and the identification of a branched intermediate, *Cell*, **75**, 1371–1377.
17. Takagi, M., Nishioka, M., Kakihara, H. et al. 1997, Characterization of DNA polymerase from *Pyrococcus* sp. strain KOD1 and its application to PCR, *Appl. Environ. Microbiol.*, **63**, 4504–4510.
18. Peler, F. B., Comb, D. G., Jack, W. E. et al. 1992, Intervening sequences in an Archaea DNA polymerase gene, *Proc. Natl. Acad. Sci. USA*, **89**, 5577–5581.
19. Niehaus, F., Frey, B., and Antranikian, G. 1997, Cloning and characterisation of a thermostable alpha-DNA polymerase from the hyperthermophilic archaeon *Thermococcus* sp. TY, *Gene*, **204**, 153–158.
20. Pietrokovski, S. 1994, Conserved sequence features of inteins (protein introns) and their use in identifying new inteins and related proteins, *Protein Science*, **3**, 2340–2350.
21. Perler, F. B., Olsen, G. J., and Adams, E. 1997, Compilation and analysis of intein sequences, *Nuc. Acids Res.*, **25**, 1087–1093.
22. Zhang, Q., Iwasaki, T., Wakagi, T., and Oshima, T. 1996, 2-Oxoacid:Ferredoxin Oxidoreductase from the Thermoacidophilic Archaeon, *Sulfolobus* sp. Strain 7, *J. Biochem.*, **120**, 587–599.







Table 5. Continued.

ORF ID	Position		Direction	Length (aa)	Predicted Product	Locus	Zscore	% Identity
	Initiation	Termination						
APE0650	438,846	436,309	<	845	phosphoenolpyruvate synthase	Swiss_Prot:P46893	230.03	62.41
APE0662	442,453	443,769	>	438	histidyl-tRNA synthetase	OWL:AB00947347	75.65	36.05
APE0665	444,837	444,124	<	237	ribose 5-phosphate isomerase	PIR:G69180	52.49	41.71
APE0669	447,540	446,854	<	228	transcriptional activator TENA	OWL:AP00000567	90.78	42.52
APE0672	449,885	448,944	<	313	malate dehydrogenase	Swiss_Prot:P49814	69.89	44.08
APE0674	452,071	453,288	<	405	aspartate aminotransferase	PIR:JC4537	65.28	34.85
APE0675	453,671	453,318	<	117	methionyl-tRNA synthetase	PIR:F70338	61.44	59.78
APE0676	454,639	453,815	<	274	3-methyl-2-oxobutanoate hydroxymethyl transferase	OWL:AP00000483	121.08	51.55
APE0680	455,882	456,316	>	144	ORF6	PIR:H70429	10.43	25.24
APE0683	457,930	458,598	>	222	phosphoserine phosphatase	OWL:AB00952815	24.95	27.89
APE0686	459,143	460,255	>	370	translation initiation factor eIF-2B alpha subunit	Swiss_Prot:Q57896	96.92	45.75
APE0689	461,506	462,813	>	435	isocitrate dehydrogenase	OWL:D86855	136.42	65.96
APE0695	465,770	464,346	<	474	Vir B11	PIR:B69292	69.46	39.67
APE0698	466,369	467,247	>	292	deoxyhypusine synthase	OWL:AP000006102	64.33	34.45
APE0703	469,708	468,551	<	385	type II DNA topoisomerase VI subunit a	Swiss_Prot:O05208	219.46	69.35
APE0706	471,436	469,739	<	565	type II DNA topoisomerase VI subunit b	Swiss_Prot:O05207	188.41	50.76
APE0708	471,778	474,039	>	753	aldehyde oxidoreductase	PIR:S73093	146.82	39.56
APE0712	476,314	475,118	<	398	N-Acylamino acid racemase	PIR:F69991	131.66	44.11
APE0715	476,453	478,012	>	519	TRK system potassium uptake protein	OWL:AB00952631	69.88	31.12
APE0739	492,098	493,135	>	345	acyl-CoA dehydrogenase	Swiss_Prot:P45857	31.92	37.95
APE0741	493,787	493,143	<	214	superoxide dismutase	Swiss_Prot:Q08713	56.35	51.69
APE0754	499,144	499,716	>	190	mitochondrial-type HSP70	PIR:B69001	66.31	41.14
APE0756	499,786	500,649	>	287	3-hydroxyisobutyrate dehydrogenase	PIR:C70303	34.50	25.89
APE0767	505,469	506,389	>	306	spermidine synthase	PIR:F70305	61.64	40.63
APE0772	508,748	509,086	>	112	signal recognition particle	PIR:E69087	28.38	42.42
APE0775	511,570	510,839	<	243	glutaredoxin-like protein	PIR:S54843	49.09	35.87
APE0777	513,100	512,300	<	266	serine/threonine protein phosphatase pp2a catalytic subunit	PIR:E69477	26.41	31.94
APE0782	515,085	516,314	>	409	N2,N2-dimethylguanosine tRNA methyltransferase	PIR:G69025	64.92	35.97
APE0789	521,170	521,580	>	136	30S ribosomal protein S8	OWL:AP00000058	60.28	51.91
APE0792	524,929	525,633	>	234	cytochrome C oxidase polypeptide II	PIR:A23711	24.20	39.83
APE0793	525,637	528,093	>	818	cytochrome C oxidase polypeptide I	OWL:MTV01256	32.29	45.91
APE0801	531,735	532,616	>	293	survival protein surE	PIR:F69367	31.74	38.28
APE0805	534,953	533,865	<	362	integrase/recombinase xerD	Swiss_Prot:P21891	41.34	39.29
APE0807	535,109	536,770	>	553	delta-1-pyrroline-5-carboxylate dehydrogenase	OWL:MTV00521	130.15	47.60
APE0809	536,898	538,352	>	484	threonyl-tRNA synthetase	PIR:F70444	65.43	42.58
APE0829	550,795	550,433	<	120	GAGE-2 protein	Swiss_Prot:Q13066	7.78	25.00
APE0872	571,101	572,450	>	449	cytosine-specific methylase	Swiss_Prot:P34905	39.49	59.34
APE0875	573,400	574,260	>	286	A/G-specific adenine glycosylase	Swiss_Prot:P29588	56.95	34.36
APE0903	593,926	594,750	>	274	alanyl-tRNA synthetase	Swiss_Prot:Q57984	40.49	35.95
APE0907	595,377	597,050	>	557	thermosome subunit	OWL:P06549	94.49	70.38
APE0912	597,205	597,987	>	260	dehydrogenase	OWL:AP000007226	83.06	47.84
APE0914	598,062	598,373	>	103	DNA-directed RNA polymerase subunit	Swiss_Prot:P36595	31.93	48.61
APE0916	600,648	599,458	<	396	mu-crystallin	OWL:AF039391	35.29	31.92
APE0917	600,515	602,017	>	500	branched-chain amino acid binding protein	PIR:F69423	71.59	32.50
APE0919	602,220	602,984	>	254	high-affinity branched-chain amino acid transport ATP-binding protein	PIR:E69423	70.28	52.82
APE0921	602,987	603,700	>	237	high-affinity branched-chain amino acid transport ATP-binding protein	PIR:D69423	57.49	45.30
APE0922	603,714	604,778	>	354	high-affinity branched-chain amino acid transport permease protein	PIR:D64458	53.78	39.51
APE0924	604,782	605,954	>	390	high-affinity branched-chain amino acid transport permease protein	PIR:E64458	68.81	33.52
APE0929	606,778	607,980	>	400	nonspecific lipid-transfer protein	PIR:C69285	161.29	58.72
APE0931	609,422	608,538	<	294	diphtine synthase	PIR:A64459	33.54	35.93
APE0935	612,995	611,982	<	337	glutamate synthase small chain	PIR:A70477	63.02	38.37
APE0941	615,984	615,181	<	267	spermidine/putrescine transport system permease protein	PIR:E69450	60.65	40.00
APE0943	616,758	615,985	<	257	spermidine/putrescine transport system permease protein	PIR:F69450	54.47	32.38
APE0944	617,719	616,742	<	325	spermidine/putrescine transport ATP-binding protein	OWL:AP00000655	43.44	55.56
APE0945	619,088	617,865	<	407	spermidine/putrescine-binding periplasmic protein	Swiss_Prot:P23861	56.10	31.40
APE0946	620,204	619,287	<	305	fumarate reductase iron-sulfur protein	PIR:B69335	80.34	51.97
APE0950	622,861	621,140	<	573	fumarate reductase flavoprotein subunit	PIR:A69335	122.06	57.14
APE0951	623,425	622,982	<	147	glycine cleavage system H protein	PIR:E70395	60.97	50.00
APE0952	623,518	624,243	>	241	hexulose-6-phosphate synthase	OWL:AB00952937	39.57	32.38
APE0955	626,217	625,330	<	295	potassium channel	PIR:S09045	29.98	49.06
APE0959	627,918	627,391	<	175	methylated-DNA--protein-cysteine methyltransferase	PIR:B69182	33.65	49.35
APE0960	629,115	627,961	<	384	p60 katanin	Swiss_Prot:Q09803	45.23	46.22
APE0976	642,023	641,706	<	105	50S ribosomal protein L14	Swiss_Prot:P54054	51.11	62.82
APE0979	644,226	642,889	<	445	amidase	OWL:AE00126911	22.39	44.19
APE0981	645,352	644,738	<	204	adenylate kinase	Swiss_Prot:P35028	77.40	44.10
APE0983	646,833	645,349	<	494	preprotein translocase secY subunit	Swiss_Prot:P49978	117.72	42.31
APE0989	649,439	649,762	>	107	nucleolar protein	OWL:AB00952250	39.76	63.16
APE0990	649,660	650,541	>	293	nucleolar protein	PIR:A69141	98.92	47.89
APE0993	650,508	651,299	>	263	purine nucleoside phosphorylase	Swiss_Prot:P50389	72.81	44.96
APE1011	659,078	659,848	>	256	protein-L-isoaspartate O-methyltransferase	PIR:B69540	86.17	55.35
APE1015	664,422	661,543	<	959	leucyl-tRNA synthetase	OWL:AP00000497	178.61	43.46
APE1023	664,590	665,474	>	294	N-acetylneuraminate lyase	Swiss_Prot:P39359	39.07	30.47
APE1026	665,554	666,162	>	202	heme exporter protein A	OWL:SC1B56	25.73	39.49
APE1038	669,592	669,963	>	123	cytidine deaminase	Swiss_Prot:P19079	60.65	54.55
APE1044	673,174	673,923	>	249	reductase	Swiss_Prot:P39605	56.59	35.11

Table 5. Continued.

ORF ID	Position		Direction	Length (aa)	Predicted Product	Locus	Zscore	% Identity
	Initiation	Termination						
APE1045	675,039	673,939	<	366	heat shock protein htpX	PIR:C69279	59.76	35.02
APE1047	675,333	676,019	>	228	translation initiation factor eIF-2B	OWL:AP000002134	23.15	29.17
APE1053	677,563	678,489	>	308	iron (III) dicitrate transport system permease protein	PIR:S74438	53.59	36.30
APE1055	678,473	679,225	>	250	iron (III) dicitrate transport ATP-binding protein	PIR:H69424	32.00	36.82
APE1061	682,347	681,316	<	343	thioredoxin reductase	OWL:AP000006131	75.05	43.71
APE1065	683,529	684,746	>	405	succinyl-CoA synthetase beta chain	Swiss_Prot:P25126	101.53	44.59
APE1072	684,753	685,646	>	297	succinyl-CoA synthetase alpha chain	PIR:A69523	98.05	62.76
APE1081	689,806	688,868	<	312	signal recognition particle protein	Swiss_Prot:P27414	88.31	51.56
APE1085	691,314	690,613	<	233	eukaryotic translation initiation factor 6	OWL:AP000002225	76.53	44.69
APE1087	691,724	691,407	<	105	50S ribosomal protein L31	Swiss_Prot:P54009	33.83	55.10
APE1090	692,832	692,371	<	153	30S ribosomal protein S19	OWL:AP00000630	86.19	58.55
APE1094	695,386	693,527	<	619	DNA ligase	Swiss_Prot:Q02093	202.86	58.79
APE1098	695,697	696,863	>	388	fmu protein	OWL:AB0094648	94.21	41.55
APE1114	702,053	701,208	<	281	methionine aminopeptidase	OWL:AP00000323	84.98	43.02
APE1119	704,785	702,971	<	604	RNase L inhibitor	OWL:AP000003259	127.34	51.08
APE1129	706,474	708,192	>	572	methionyl-tRNA synthetase	OWL:AP000004125	122.87	42.22
APE1132	708,253	708,612	>	119	30S ribosomal protein S24	PIR:A69389	30.99	38.54
APE1135	708,868	709,917	>	349	O-sialoglycoprotein endopeptidase	OWL:AB00946412	109.50	47.73
APE1139	711,198	711,650	>	150	30S ribosomal protein S13	Swiss_Prot:P54012	86.08	60.42
APE1140	711,805	713,226	>	473	aspartate kinase	Swiss_Prot:Q57991	63.12	39.02
APE1143	713,223	714,272	>	349	aspartate-semialdehyde dehydrogenase	PIR:E69206	108.95	47.09
APE1144	714,263	715,267	>	334	homoserine dehydrogenase	PIR:C69154	30.80	33.59
APE1154	719,058	719,723	>	221	30S ribosomal protein S3	OWL:AB00946620	91.91	50.79
APE1178	730,712	730,149	<	187	dTDP-4-dehydrothymine 3,5-epimerase	OWL:LIU61226	66.16	47.28
APE1179	731,633	730,716	<	305	dTDP-4-dehydrothymine reductase	OWL:AP000002111	62.86	41.25
APE1180	732,619	731,627	<	330	dTDP-glucose 4,6-dehydratase	PIR:D69290	77.27	50.75
APE1181	733,699	732,632	<	355	glucose-1-phosphate thymidyltransferase	PIR:E69290	78.50	39.39
APE1191	746,786	745,695	<	363	capM protein	PIR:C69098	43.66	30.73
APE1192	747,663	746,791	<	290	dolichol monophosphate mannose synthase	PIR:E69322	19.29	28.49
APE1193	748,343	747,756	<	195	capsular synthesis regulator component B	OWL:AF0712151	11.39	22.68
APE1195	749,434	748,874	<	186	adenylylsulfate kinase	PIR:S74917	62.65	45.24
APE1197	750,622	751,791	>	389	sulfate adenylyltransferase	PIR:B69839	118.71	48.54
APE1202	753,560	754,642	>	360	sugar phosphate transferase	PIR:D64437	76.76	37.91
APE1205	754,793	756,670	>	625	glucosamine--fructose-6-phosphate aminotransferase	OWL:AB009473	103.64	42.42
APE1208	756,722	757,804	>	360	glucose-1-phosphate thymidyltransferase	PIR:D64437	22.09	31.85
APE1211	763,287	763,670	>	127	gastrin precursor	Swiss_Prot:P48757	9.35	25.97
APE1221	772,391	773,461	>	356	tropomyosin	OWL:U9387282	29.36	42.50
APE1223	773,599	774,780	>	393	cystathionine beta-synthase	PIR:C42790	61.64	41.40
APE1226	774,777	775,931	>	384	cystathionine gamma-lyase	Swiss_Prot:P18757	89.33	52.37
APE1245	792,718	791,720	<	332	alcohol dehydrogenase	Swiss_Prot:Q57517	44.29	29.32
APE1253	795,518	796,438	>	306	ABC-transporter ATP-binding protein	PIR:B69377	43.82	46.67
APE1255	796,450	797,211	>	253	ABC transporter ATP-binding protein	OWL:AP000006241	46.27	41.18
APE1261	800,410	800,093	<	105	formylmethanofuran dehydrogenase	OWL:MKFWUGDB6	8.17	33.33
APE1265	801,213	802,235	>	340	dehydrogenase	OWL:AP000002294	36.50	31.48
APE1275	806,006	807,409	>	467	tryptophanase	Swiss_Prot:P31014	122.59	49.02
APE1279	810,828	810,304	<	174	riboflavin biosynthesis enzyme	Swiss_Prot:Q58085	37.56	44.36
APE1285	813,976	815,097	>	373	acetylpolymine aminohydrolase	OWL:AP000005176	56.34	36.69
APE1288	816,917	820,699	>	1,260	nitrate reductase alpha chain	Swiss_Prot:P42175	176.97	51.32
APE1294	820,733	822,133	>	466	nitrate reductase beta chain	Swiss_Prot:P42176	128.61	50.32
APE1300	824,070	824,618	>	182	nitrate reductase delta chain	Swiss_Prot:P42178	20.90	32.30
APE1301	825,361	824,615	<	248	high-affinity branched-chain amino acid transport ATP-binding protein	Swiss_Prot:P30294	48.66	48.00
APE1304	827,831	826,782	<	349	high-affinity branched-chain amino acid transport permease protein	PIR:H69352	42.39	32.66
APE1305	828,720	827,836	<	294	high-affinity branched-chain amino acid transport permease protein	PIR:A69353	43.40	29.97
APE1307	830,682	828,727	<	651	long-chain-fatty-acid--CoA ligase	Swiss_Prot:P44446	75.18	27.30
APE1308	831,438	830,686	<	250	high-affinity branched-chain amino acid transport ATP-binding protein	PIR:G69352	56.40	44.00
APE1310	831,654	833,243	>	529	luciferase	PIR:A69831	49.31	28.60
APE1317	834,911	835,831	>	306	methanol dehydrogenase regulator	PIR:C69791	104.86	50.17
APE1324	839,616	839,092	<	174	heterocyst differentiation protein	PIR:S58288	20.15	27.70
APE1328	840,722	843,079	>	785	hyuA	PIR:D64420	208.87	49.63
APE1331	843,086	844,774	>	562	hyuB	PIR:C64420	181.47	52.01
APE1335	845,237	846,256	>	339	deblocking aminopeptidase	OWL:AP000002215	107.81	45.81
APE1340	849,459	853,589	>	1,376	reverse gyrase	OWL:PFU66557	177.77	41.88
APE1346	857,515	855,533	<	660	acetyl-CoA synthetase	PIR:G69371	161.31	54.10
APE1350	858,472	858,005	<	155	molybdenum cofactor biosynthesis protein	PIR:F69518	72.97	53.90
APE1353	858,753	862,574	>	1,273	ATP-dependent DNA helicase	PIR:F69901	25.34	36.81
APE1366	869,385	868,912	<	157	6,7-dimethyl-8-ribityllumazine synthase	PIR:H69515	82.32	57.04
APE1367	869,642	871,822	>	726	transitional endoplasmic reticulum ATPase	Swiss_Prot:Q07590	110.09	70.29
APE1376	874,765	878,433	>	1,222	reverse gyrase	OWL:PFU66557	164.08	36.83
APE1380	878,464	879,414	>	316	transport ATP-binding protein	OWL:AP00000671	32.74	36.04
APE1382	880,421	881,773	>	450	hyuE	OWL:SC1A65	20.17	23.87
APE1386	882,894	884,165	>	423	glutamate dehydrogenase	Swiss_Prot:P80053	97.17	56.79
APE1395	888,205	889,215	>	336	dipeptide transport system permease protein	OWL:AP000006115	132.26	52.71
APE1396	889,217	890,380	>	387	dipeptide transport system permease protein	OWL:AP000006116	152.98	56.28
APE1399	890,388	891,389	>	333	dipeptide transport system ATP-binding protein	OWL:AP000006117	103.01	66.98
APE1402	891,395	892,438	>	347	oligopeptide transport ATP-binding protein	OWL:AP000006118	78.84	65.20









Table 5. Continued.

ORF ID	Position		Direction	Length (aa)	Predicted Product	Locus	Zscore	% Identity
	Initiation	Termination						
APES008	196,429	196,611	<	60	repressor protein	Swiss_Prot:P06965	8.71	32.65
APES015	306,121	306,342	<	73	30S ribosomal protein S27	Swiss_Prot:P54028	42.73	52.54
APES016	306,359	306,550	<	63	50S ribosomal protein L44	Swiss_Prot:P27074	23.68	49.12
APES019	315,792	315,986	<	64	ribonuclease p3	PIR:S69085	18.00	87.50
APES022	361,519	361,752	>	77	small nuclear ribonucleoprotein	PIR:C69186	36.33	46.67
APES023	361,782	361,970	>	62	50S ribosomal protein L37	PIR:C71028	53.31	62.90
APES024	370,582	370,857	>	91	50S ribosomal protein L13	OWL:RLAJ3198	18.41	49.06
APES041	658,128	658,400	>	90	DNA-directed RNA polymerase, subunit M	Swiss_Prot:O13896	8.62	33.96
APES045	760,865	761,038	>	57	aerolysin regulatory protein	Swiss_Prot:P09165	10.52	39.47
APES054	917,209	917,469	<	86	50S ribosomal protein L37	PIR:B71204	42.89	46.91
APES058	1,103,406	1,103,639	>	77	DNA-directed RNA polymerase, subunit N	PIR:A69391	41.43	45.76
APES061	1,175,870	1,176,142	<	90	DNA-directed RNA polymerase, subunit H	PIR:D69485	51.10	58.57
APES067	1,305,985	1,306,209	<	74	30S ribosomal protein S17	Swiss_Prot:P54026	44.93	58.93
APES068	1,395,524	1,395,676	>	50	30S ribosomal protein S30	Swiss_Prot:Q05472	19.30	46.51
APES070	1,513,640	1,513,831	<	63	50S ribosomal protein L24	PIR:F69345	48.89	60.38
APES071	1,513,891	1,514,166	>	91	30S ribosomal protein S28	PIR:C71025	68.57	76.81
APES074	1,657,629	1,657,865	<	78	Acetyl-CoA synthetase	Swiss_Prot:P27097	20.50	53.66

The ORF name, initiation and termination positions, direction of reading and length of amino acid residues of ORFs are shown in the left five columns. The right four columns represent the probable gene products, gene locus of genes which showed the highest similarity to corresponding ORFs, Zscore taken from Smith-Waterman similarity search and percent identity between the reported genes and assigned ORFs.



Table 6. List of ORFs with motifs.

ORF ID	Position		Direction	Length (aa)	Motifs assigned
	Initiation	Termination			
APE0004	2,174	1,260	<	304	gram-positive cocci surface proteins 'anchoring' hexapeptide
APE0033	20,898	22,379	>	493	ATP/GTP-binding site motif A (P-loop)
APE0034	22,033	21,668	<	121	prokaryotic membrane lipoprotein lipid attachment site
APE0054	33,600	33,058	<	180	tonB-dependent receptor proteins signatures
APE0073	50,533	52,197	>	554	N-6 Adenine-specific DNA methylases signature
APE0080	56,596	54,512	<	694	ATP/GTP-binding site motif A (P-loop)
APE0090	64,466	63,975	<	163	prokaryotic membrane lipoprotein lipid attachment site
APE0111	81,414	81,947	>	177	prokaryotic membrane lipoprotein lipid attachment site
APE0112	83,240	83,749	>	169	prokaryotic membrane lipoprotein lipid attachment site
APE0130	97,001	97,999	>	332	prokaryotic membrane lipoprotein lipid attachment site
APE0138	103,068	102,703	<	121	ATP synthase alpha and beta subunits signature
APE0154	115,851	114,493	<	452	prokaryotic membrane lipoprotein lipid attachment site
APE0163	124,215	122,824	<	463	prokaryotic membrane lipoprotein lipid attachment site
APE0170	126,927	127,283	>	118	ATP/GTP-binding site motif A (P-loop) ; prokaryotic membrane lipoprotein lipid attachment site
APE0179	133,085	132,240	<	281	SRP54-type proteins GTP-binding domain signature
APE0186	137,641	137,309	<	110	crystallins beta and gamma 'Greek key' motif signature
APE0189	139,900	139,598	<	100	prokaryotic membrane lipoprotein lipid attachment site
APE0219	159,880	158,495	<	461	ATP/GTP-binding site motif A (P-loop)
APE0223	161,921	162,328	>	135	prokaryotic membrane lipoprotein lipid attachment site
APE0232	167,920	168,267	>	115	prokaryotic membrane lipoprotein lipid attachment site
APE0242	175,106	175,429	>	107	2Fe-2S ferredoxins, iron-sulfur binding region signature
APE0247	177,017	177,601	>	194	thiolases signatures
APE0256	183,717	183,160	<	185	prokaryotic membrane lipoprotein lipid attachment site
APE0283	206,831	208,840	>	669	ATP/GTP-binding site motif A (P-loop)
APE0292	213,021	212,701	<	106	gram-positive cocci surface proteins 'anchoring' hexapeptide
APE0294	213,536	214,009	>	157	multicopper oxidases signatures
APE0298	215,194	215,586	>	130	prokaryotic membrane lipoprotein lipid attachment site
APE0305	221,950	222,600	>	216	prokaryotic membrane lipoprotein lipid attachment site
APE0340	249,296	247,047	<	749	ATP/GTP-binding site motif A (P-loop)
APE0366	259,045	259,611	>	188	prokaryotic membrane lipoprotein lipid attachment site
APE0386	270,592	271,005	>	137	prokaryotic membrane lipoprotein lipid attachment site
APE0397	274,689	275,045	>	118	prokaryotic membrane lipoprotein lipid attachment site
APE0409	282,320	281,976	<	114	serine proteases, trypsin family, active sites
APE0415	291,639	288,073	<	1,188	regulator of chromosome condensation (RCC1) signatures
APE0424	298,856	299,269	>	137	prokaryotic membrane lipoprotein lipid attachment site
APE0428	303,153	301,243	<	636	prokaryotic membrane lipoprotein lipid attachment site
APE0431	303,833	304,396	>	187	ATP/GTP-binding site motif A (P-loop)
APE0434	305,005	305,367	>	120	prokaryotic membrane lipoprotein lipid attachment site
APE0472	329,651	329,133	<	172	tonB-dependent receptor proteins signatures
APE0481	335,324	335,016	<	102	4Fe-4S ferredoxins, iron-sulfur binding region signature
APE0486	338,138	337,794	<	114	prokaryotic membrane lipoprotein lipid attachment site
APE0540	372,037	370,874	<	387	ATP/GTP-binding site motif A (P-loop)
APE0548	376,327	375,746	<	193	prokaryotic membrane lipoprotein lipid attachment site
APE0554	378,497	379,144	>	215	N-6 Adenine-specific DNA methylases signature
APE0565	384,956	385,537	>	193	prokaryotic membrane lipoprotein lipid attachment site
APE0576	388,764	389,156	>	130	prokaryotic membrane lipoprotein lipid attachment site
APE0588	393,438	396,176	>	912	prokaryotic membrane lipoprotein lipid attachment site
APE0614	418,219	416,585	<	544	lipocalin signature
APE0617	418,227	418,640	>	137	prokaryotic membrane lipoprotein lipid attachment site
APE0620	425,424	419,623	<	1,933	mitochondrial energy transfer proteins signature
APE0625	427,791	427,396	<	131	prokaryotic membrane lipoprotein lipid attachment site
APE0626	428,224	427,715	<	169	prokaryotic membrane lipoprotein lipid attachment site
APE0627	429,153	428,758	<	131	zinc finger, C2H2 type, domain
APE0641	434,604	433,555	<	349	thioredoxin family active site
APE0643	434,144	434,674	>	176	prokaryotic membrane lipoprotein lipid attachment site
APE0653	438,163	438,798	>	211	gram-positive cocci surface proteins 'anchoring' hexapeptide
APE0705	469,437	469,916	>	159	gram-positive cocci surface proteins 'anchoring' hexapeptide; prokaryotic membrane lipoprotein lipid attachment site
APE0759	500,630	501,706	>	358	gram-positive cocci surface proteins 'anchoring' hexapeptide
APE0774	509,651	510,844	>	397	N-6 Adenine-specific DNA methylases signature
APE0787	519,490	520,416	>	308	transcription factor TFIIIB repeat signature
APE0796	528,664	529,365	>	233	legume lectins signatures
APE0823	545,622	546,302	>	226	prenyl group binding site (CAAX box) ; prokaryotic membrane lipoprotein lipid attachment site
APE0824	546,649	545,948	<	233	neutral zinc metalloproteinases, zinc-binding region signature
APE0837	553,499	554,215	>	238	aminoacyl-transfer RNA synthetases class-II signatures
APE0851	559,911	559,507	<	134	prokaryotic membrane lipoprotein lipid attachment site
APE0867	567,267	568,112	>	281	neutral zinc metalloproteinases, zinc-binding region signature
APE0880	576,234	577,838	>	534	ATP/GTP-binding site motif A (P-loop)
APE0892	586,334	587,476	>	380	ATP/GTP-binding site motif A (P-loop)
APE0897	588,901	589,251	>	116	prokaryotic membrane lipoprotein lipid attachment site
APE0898	589,753	592,311	>	852	gram-positive cocci surface proteins 'anchoring' hexapeptide
APE0963	629,175	629,762	>	195	prokaryotic membrane lipoprotein lipid attachment site
APE0984	646,204	645,629	<	191	prokaryotic membrane lipoprotein lipid attachment site
APE0999	652,949	653,536	>	195	ATP/GTP-binding site motif A (P-loop)
APE1007	657,169	658,074	>	301	aminoacyl-transfer RNA synthetases class-II signatures
APE1076	686,760	686,443	<	105	aldehyde dehydrogenases active sites
APE1082	689,187	689,591	>	134	prokaryotic membrane lipoprotein lipid attachment site
APE1088	691,810	692,289	>	159	prokaryotic membrane lipoprotein lipid attachment site
APE1122	704,552	704,968	>	138	binding-protein-dependent transport systems inner membrane components signature

Table 6. Continued.

ORF ID	Position		Direction	Length (aa)	Motifs assigned
	Initiation	Termination			
APE1151	718,419	717,868	<	183	prokaryotic membrane lipoprotein lipid attachment site
APE1175	728,323	727,871	<	150	prokaryotic membrane lipoprotein lipid attachment site
APE1213	764,158	768,282	>	1,374	serine proteases, subtilase family, active sites
APE1216	768,368	769,969	>	533	ribosomal protein S2 signatures
APE1218	770,849	770,142	<	235	prokaryotic membrane lipoprotein lipid attachment site
APE1234	781,349	780,435	<	304	prokaryotic membrane lipoprotein lipid attachment site
APE1244	791,512	789,800	<	570	lipases, serine active site ; prokaryotic membrane lipoprotein lipid attachment site
APE1274	804,971	805,525	>	184	prokaryotic membrane lipoprotein lipid attachment site
APE1277	809,886	809,320	<	188	phosphopantetheine attachment site
APE1303	826,697	825,405	<	430	prokaryotic membrane lipoprotein lipid attachment site
APE1309	830,771	831,136	>	121	prokaryotic membrane lipoprotein lipid attachment site
APE1311	832,234	831,854	<	126	bacterial regulatory proteins, lysR family signature
APE1315	833,830	834,936	>	368	crystallins beta and gamma 'Greek key' motif signature
APE1322	837,325	837,822	>	165	aminoacyl-transfer RNA synthetases class-II signatures
APE1326	840,888	839,725	<	387	aldehyde dehydrogenases active sites; nitrite and sulfite reductases iron-sulfur/siroheme-binding site; prokaryotic membrane lipoprotein lipid attachment site
APE1348	856,641	857,072	>	143	prokaryotic membrane lipoprotein lipid attachment site
APE1392	887,352	886,810	<	180	prokaryotic membrane lipoprotein lipid attachment site
APE1413	901,702	902,028	>	108	prokaryotic membrane lipoprotein lipid attachment site
APE1433	909,998	910,468	>	156	prokaryotic membrane lipoprotein lipid attachment site
APE1435	911,306	911,803	>	165	prokaryotic membrane lipoprotein lipid attachment site
APE1439	913,927	914,475	>	182	prokaryotic membrane lipoprotein lipid attachment site
APE1468	933,223	932,807	<	138	ATP synthase alpha and beta subunits signature
APE1482	941,320	940,946	<	124	prokaryotic membrane lipoprotein lipid attachment site
APE1493	946,525	947,247	>	240	prokaryotic membrane lipoprotein lipid attachment site
APE1499	950,583	951,041	>	152	prokaryotic membrane lipoprotein lipid attachment site
APE1504	953,346	953,038	<	102	prokaryotic membrane lipoprotein lipid attachment site
APE1519	960,138	960,536	>	132	prokaryotic membrane lipoprotein lipid attachment site
APE1534	969,058	968,192	<	288	prokaryotic membrane lipoprotein lipid attachment site
APE1544	973,098	974,015	>	305	prokaryotic membrane lipoprotein lipid attachment site
APE1553	980,168	980,536	>	122	prokaryotic membrane lipoprotein lipid attachment site
APE1561	987,536	987,952	>	138	neutral zinc metallopeptidases, zinc-binding region signature
APE1566	988,955	989,299	>	114	prokaryotic membrane lipoprotein lipid attachment site
APE1615	1,024,726	1,023,671	<	351	TNF family signature and profile
APE1619	1,027,311	1,027,715	>	134	prokaryotic membrane lipoprotein lipid attachment site
APE1709	1,076,164	1,076,556	>	130	prokaryotic membrane lipoprotein lipid attachment site
APE1711	1,078,508	1,079,467	>	319	aldo/keto reductase family signatures
APE1715	1,081,752	1,080,844	<	302	peroxidases signatures
APE1716	1,081,808	1,082,122	>	104	glycine radical signature
APE1719	1,085,123	1,085,707	>	194	cytochrome c family heme-binding site signature
APE1726	1,092,371	1,091,922	<	149	prokaryotic membrane lipoprotein lipid attachment site
APE1733	1,097,503	1,097,066	<	145	prokaryotic membrane lipoprotein lipid attachment site
APE1748	1,103,355	1,102,810	<	181	prokaryotic membrane lipoprotein lipid attachment site
APE1757	1,107,397	1,108,080	>	227	aldo/keto reductase family signatures
APE1787	1,124,639	1,124,995	>	118	prokaryotic membrane lipoprotein lipid attachment site
APE1799	1,133,202	1,134,035	>	277	cytochrome c family heme-binding site signature
APE1802	1,134,791	1,135,135	>	114	prokaryotic membrane lipoprotein lipid attachment site
APE1810	1,140,809	1,142,485	>	558	ATP/GTP-binding site motif A (P-loop)
APE1812	1,141,995	1,141,384	<	203	bacterial regulatory proteins, gntR family signature
APE1815	1,145,037	1,144,132	<	301	prokaryotic membrane lipoprotein lipid attachment site
APE1820	1,149,068	1,147,656	<	470	apoptosis regulator proteins, Bcl-2 family BH domains signatures/profiles
APE1855	1,172,045	1,172,368	>	107	prokaryotic membrane lipoprotein lipid attachment site
APE1860	1,174,574	1,175,329	>	251	prokaryotic membrane lipoprotein lipid attachment site
APE1879	1,191,036	1,190,704	<	110	prokaryotic membrane lipoprotein lipid attachment site
APE1883	1,191,742	1,192,152	>	136	prokaryotic membrane lipoprotein lipid attachment site
APE1920	1,214,934	1,215,437	>	167	prokaryotic membrane lipoprotein lipid attachment site
APE1923	1,216,957	1,216,640	<	105	prokaryotic membrane lipoprotein lipid attachment site
APE1924	1,217,925	1,218,350	>	141	prokaryotic membrane lipoprotein lipid attachment site
APE1956	1,236,638	1,235,436	<	400	lipocalin signature
APE1973	1,248,412	1,250,532	>	706	ribosomal protein S2 signatures
APE1978	1,252,953	1,253,483	>	176	prokaryotic membrane lipoprotein lipid attachment site
APE2001	1,264,439	1,263,975	<	154	bacterial regulatory proteins, gntR family signature
APE2005	1,265,616	1,266,005	<	129	prokaryotic membrane lipoprotein lipid attachment site
APE2008	1,266,696	1,267,109	>	137	prokaryotic membrane lipoprotein lipid attachment site
APE2028	1,278,621	1,278,151	<	156	trp-Asp (WD-40) repeats signature
APE2040	1,284,563	1,284,210	<	117	prokaryotic membrane lipoprotein lipid attachment site
APE2043	1,285,931	1,285,602	<	109	ribosomal protein S14 signature
APE2056	1,292,756	1,292,289	<	155	mut T domain signature
APE2079	1,313,114	1,312,668	<	148	prokaryotic membrane lipoprotein lipid attachment site
APE2082	1,314,444	1,313,980	<	154	prokaryotic membrane lipoprotein lipid attachment site
APE2083	1,314,454	1,314,062	<	130	immunoglobulins and major histocompatibility complex proteins signature
APE2093	1,321,205	1,322,749	>	514	ATP/GTP-binding site motif A (P-loop)
APE2119	1,338,235	1,336,766	<	489	prokaryotic membrane lipoprotein lipid attachment site
APE2123	1,339,800	1,340,186	>	128	prokaryotic membrane lipoprotein lipid attachment site
APE2129	1,345,980	1,346,351	>	123	gram-positive cocci surface proteins 'anchoring' hexapeptide
APE2132	1,346,786	1,347,175	>	129	prokaryotic membrane lipoprotein lipid attachment site
APE2138	1,351,675	1,351,307	<	122	prokaryotic membrane lipoprotein lipid attachment site
APE2148	1,359,442	1,359,897	>	151	prokaryotic membrane lipoprotein lipid attachment site
APE2163	1,373,730	1,374,053	>	107	peroxidases signatures

Table 6. Continued.

ORF ID	Position		Direction	Length (aa)	Motifs assigned
	Initiation	Termination			
APE2164	1,374,060	1,374,467	>	135	ATP/GTP-binding site motif A (P-loop)
APE2169	1,377,928	1,378,284	>	118	prokaryotic membrane lipoprotein lipid attachment site
APE2187	1,386,739	1,385,831	<	302	prokaryotic membrane lipoprotein lipid attachment site
APE2190	1,389,893	1,390,957	>	354	EF-hand calcium-binding domain ; gram-positive cocci surface proteins 'anchoring' hexapeptide
APE2191	1,390,932	1,390,258	<	224	prokaryotic membrane lipoprotein lipid attachment site
APE2211	1,403,501	1,403,055	<	148	prokaryotic membrane lipoprotein lipid attachment site
APE2233	1,414,987	1,414,583	<	134	prokaryotic membrane lipoprotein lipid attachment site
APE2246	1,422,983	1,423,549	>	188	eukaryotic putative RNA-binding region RNP-1 signature
APE2292	1,467,563	1,466,016	<	515	lipases, serine active site ; tonB-dependent receptor proteins signatures
APE2310	1,477,855	1,478,952	>	365	prokaryotic membrane lipoprotein lipid attachment site
APE2325	1,486,623	1,487,096	>	157	phospholipase A2 active sites signatures
APE2329	1,489,289	1,488,822	<	155	prokaryotic membrane lipoprotein lipid attachment site
APE2358	1,504,387	1,503,749	<	212	prokaryotic membrane lipoprotein lipid attachment site
APE2368	1,508,742	1,509,155	>	137	pfkB family of carbohydrate kinases signatures
APE2383	1,516,744	1,516,247	<	165	ATP synthase alpha and beta subunits signature
APE2385	1,517,802	1,517,449	<	117	prokaryotic membrane lipoprotein lipid attachment site
APE2401	1,523,649	1,524,407	>	252	prokaryotic membrane lipoprotein lipid attachment site
APE2405	1,525,109	1,524,783	<	108	gram-positive cocci surface proteins 'anchoring' hexapeptide
APE2409	1,526,992	1,526,483	<	169	ATP/GTP-binding site motif A (P-loop) ; prokaryotic membrane lipoprotein lipid attachment site
APE2411	1,527,536	1,526,943	<	197	G-protein coupled receptors signature
APE2429	1,538,592	1,539,062	>	156	prokaryotic membrane lipoprotein lipid attachment site
APE2434	1,540,709	1,540,353	<	118	prokaryotic membrane lipoprotein lipid attachment site
APE2454	1,558,082	1,557,762	<	106	prokaryotic membrane lipoprotein lipid attachment site
APE2478	1,575,425	1,575,826	>	133	prokaryotic membrane lipoprotein lipid attachment site
APE2487	1,579,230	1,578,730	<	166	prokaryotic membrane lipoprotein lipid attachment site
APE2508	1,593,077	1,592,763	<	104	prokaryotic membrane lipoprotein lipid attachment site
APE2534	1,610,413	1,610,099	<	104	prokaryotic membrane lipoprotein lipid attachment site
APE2552	1,624,302	1,623,898	<	134	ATP/GTP-binding site motif A (P-loop) ; prokaryotic membrane lipoprotein lipid attachment site
APE2560	1,628,248	1,628,625	>	125	prokaryotic membrane lipoprotein lipid attachment site
APE2567	1,631,425	1,630,772	<	217	prokaryotic membrane lipoprotein lipid attachment site
APE2568	1,632,097	1,631,432	<	221	prokaryotic membrane lipoprotein lipid attachment site
APE2593	1,649,757	1,650,080	>	107	prokaryotic membrane lipoprotein lipid attachment site
APE2600	1,652,475	1,652,804	>	109	prokaryotic membrane lipoprotein lipid attachment site
APE2614	1,664,255	1,663,881	<	124	prokaryotic membrane lipoprotein lipid attachment site
APES013	295,342	295,518	>	58	prokaryotic membrane lipoprotein lipid attachment site
APES018	309,642	309,842	<	66	lipocalin signature
APES034	503,403	503,633	>	76	prokaryotic membrane lipoprotein lipid attachment site
APES036	549,567	549,731	>	54	phospholipase A2 active sites signatures
APES037	568,627	568,878	>	83	prokaryotic membrane lipoprotein lipid attachment site
APES042	729,946	730,107	>	53	prenyl group binding site (CAAX box)
APES048	837,843	838,043	>	66	ATP/GTP-binding site motif A (P-loop)
APES051	869,419	869,589	<	56	prokaryotic membrane lipoprotein lipid attachment site
APES052	871,917	872,078	<	53	prenyl group binding site (CAAX box)
APES059	1,103,409	1,103,606	<	65	prokaryotic membrane lipoprotein lipid attachment site
APES072	1,563,366	1,563,638	>	90	prokaryotic membrane lipoprotein lipid attachment site
APES073	1,614,964	1,615,182	>	72	zinc carboxypeptidases, zinc-binding regions signatures

The ORF name, initiation and termination positions, direction of reading and length of amino acid residues of ORFs are shown in the left five columns, and the assigned protein-motifs in the right column.

