Protein Science (1994), *3*:1329–1340. Cambridge University Press. Printed in the USA.
Copyright © 1994 The Protein Society

The sequence of a subtilisin-type protease (aerolysin) from the hyperthermophilic archaeum *Pyrobaculum aerophilum* reveals sites important to thermostability

PAUL VÖLKL,[1,2] PETER MARKIEWICZ,[1] KARL O. STETTER,[2]
AND JEFFREY H. MILLER[1]

[1] Department of Microbiology and Molecular Genetics and the Molecular Biology Institute, University of California, Los Angeles, California 90024
[2] Archaeenzentrum, Universität Regensburg, 93053 Regensburg, Germany

(RECEIVED December 16, 1994; ACCEPTED April 20, 1994)

Abstract

The hyperthermophilic archaeum *Pyrobaculum aerophilum* grows optimally at 100 °C and pH 7.0. Cell homogenates exhibit strong proteolytic activity within a temperature range of 80–130 °C. During an analysis of cDNA and genomic sequence tags, a genomic clone was recovered showing strong sequence homology to alkaline subtilisins of *Bacillus* sp. The total DNA sequence of the gene encoding the protease (named "aerolysin") was determined. Multiple sequence alignment with 15 different serine-type proteases showed greatest homology with subtilisins from gram-positive bacteria rather than archaeal or eukaryal serine proteases.

Models of secondary and tertiary structure based on sequence alignments and the tertiary structures of subtilisin Carlsberg, BPN′, thermitase, and protease K were generated for *P. aerophilum* subtilisin. This allowed identification of sites potentially contributing to the thermostability of the protein. One common transition put alanines at the beginning and end of surface alpha-helices. Aspartic acids were found at the N-terminus of several surface helices, possibly increasing stability by interacting with the helix dipole. Several of the substitutions in regions expected to form surface loops were adjacent to each other in the tertiary structure model.

Keywords: modeling; *Pyrobaculum*; subtilisin; thermophile; thermostability

Hyperthermophilic organisms are of great theoretical and commercial interest. These organisms are distinct from "classical" thermophiles such as *Thermus aquaticus* in having much higher temperature optima for growth (between 100 and 115 °C). The majority of the Archaea (formerly Archaebacteria) are hyperthermophilic, and some grow at temperatures above 110 °C (Woese et al., 1990; Stetter, 1993).

Analysis of the proteins of these organisms is expected to be particularly revealing. It is believed that the progenote, the last common ancestor of all life, was probably a hyperthermophile (Woese, 1990). Hyperthermophiles may therefore provide a window into the requirements for protein thermostability. Isolated proteins from these organisms are frequently active at temperatures exceeding 150 °C (Schuliger et al., 1993), and the phylogenetic tree indicates that such extreme thermostability may be a primitive, not an acquired, characteristic. If mesophilic proteins are cold-active mutants of their original form, studies of hyperthermophilic proteins may be necessary to explain protein stability.

Unfortunately, cultivation of hyperthermophilic organisms is difficult because of their high growth temperatures, low growth titers, strict anaerobiosis, sulfur dependence, and other features. For this reason, most of the exploratory work has been done using proteins from *Thermus* sp. with much lower temperature optima (65 °C). Recently, a novel hyperthermophilic archaeum, *Pyrobaculum aerophilum* (type strain: IM2), was described. This organism grows optimally at 100 °C and pH 7.0 (Völkl et al., 1993). It is the only facultatively aerobic member of the slowly evolving order Thermoproteales, and it has several features that make it a suitable candidate for developing genetic systems. To begin analysis of the genome of *P. aerophilum*, we collected sequence information from randomly chosen clones of genomic and cDNA libraries following the sequence tag methodology (Adams et al., 1993). The characteristics and applications of the sequence tags of *P. aerophilum* will be reported elsewhere (Völkl et al., manuscript in prep.). In the following, we describe the analysis of a gene from *P. aerophilum* showing strong similarities to *Bacillus*-type subtilisins. Subtilisin is one of the best stud-

Reprint requests to: Jeffrey H. Miller, Molecular Biology Institute, University of California, 405 Hilgard Avenue, Los Angeles, California 90024; e-mail: jmiller@inherit2.dna.ucla.edu.

ied proteins (Wells et al., 1987). The 3-dimensional structure of several variants is known, and sequences of related serine proteases are available from different organisms. It is of commercial value, being used as an additive in detergents, and several studies have attempted to improve its thermostability (Mitchinson & Wells, 1989; Eijsink et al., 1992) by protein engineering. The analysis of a hyperthermophilic subtilisin should improve our understanding of protein stability at high temperatures.

To this end, the complete sequence of the *P. aerophilum* subtilisin gene was determined and its amino acid sequence compared with other subtilisins. Preliminary analysis of the sequence allows us to draw conclusions about the structure, thermostability, and evolution of this protein.

Results

Cloning and sequencing of a serine-type protease from P. aerophilum

Randomly chosen clones from a 1–2-kb genomic DNA library and a poly A primed cDNA library of *P. aerophilum* were sequenced (Völkl et al., manuscript in prep.). One of the genomic clones (GSPA-35) carried a 1-kb insert with an open reading frame (ORF) matching the N-terminal sequence of various *Bacillus* sp. subtilisins.

To clone the remainder of the gene, chromosomal DNA of *P. aerophilum* was digested with restriction endonucleases and probed by Southern blot hybridization with a radiolabeled 18-mer oligonucleotide derived from GSPA-35. Fragments hybridizing to the probe were recovered and subcloned (Fig. 1). Additional plaque lift hybridizations recovered several clones from the poly A cDNA library containing the subtilisin gene. The complete sequence is given in Figure 2. Comparison of the cDNA and genomic sequences showed no introns were present, despite a report of an intervening sequence in *P. aerophilum* 16S rRNA (Burggraf et al., 1993). Some cDNA clones started well upstream of the predicted N-terminus. This, and the fact that primer extension experiments gave signals 74 and 400 bases upstream of the ORF, indicated that transcription may proceed from an upstream promoter. Sequencing of other cDNAs from *P. aerophilum* has shown that many transcripts are polycistronic (Völkl et al., manuscript in prep.).

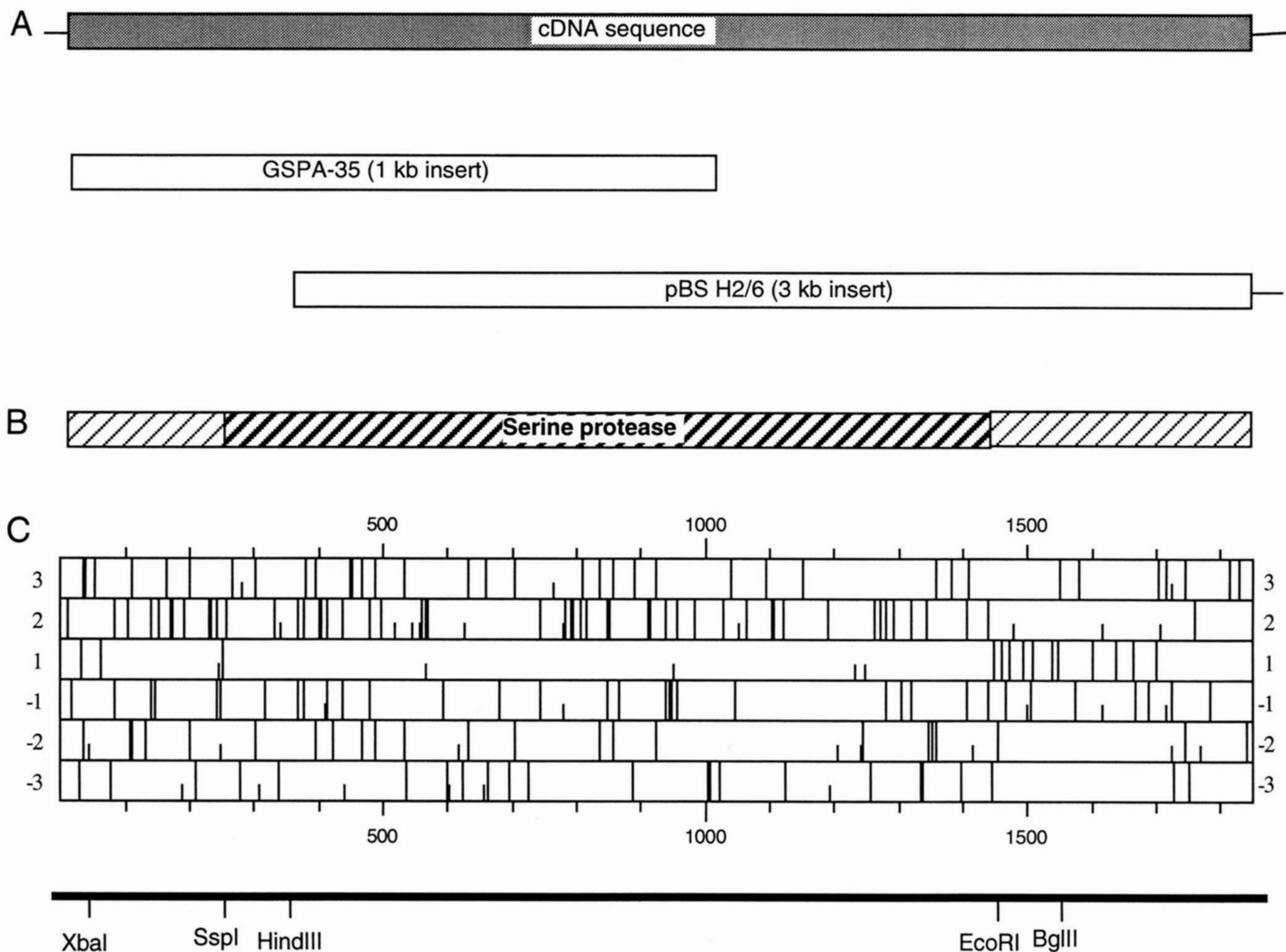


Fig. 1. Schematic diagram of a serine-type protease gene from *P. aerophilum*. **A:** Clones used to construct the total sequence containing the protease gene. ESUB3, cDNA clone obtained by plaque lift hybridization of poly A primed cDNA library; GSPA-35, randomly sequenced clone of genomic 1–2-kb library; pBS H2/6, *Hin*d III fragment obtained by Southern blot and colony lift hybridization. **B:** The open reading frame (+1) encoding the serine-type protease. **C:** Restriction map.

```
GGAACAAAGCTGAGCTCACgGTGCGCGCTCTAGACTAGTGATCCAtTGGCGAGTGACTTGTGAATACTCCAAGCGCtTTAC
-162
TTAATCCAGTGGGAGGGCAAGCTGACTATTAGACAAGCCCCCCAGTACTTCAACGAGTTAGATTTACAATTGAGAATCGGC

GCTGAGGTGATAGAAACTGCGAAAAGCATAGGCGTTTCTAAAAAAGTTCAAAAAAGAGTTTCCGCTGTAATAGACGAATTA
1/1                                     31/11
ATG TCA TAA TAT TTG AAA AGA CGT AAA AAA CTG GCT CAA TGC CAA AGG TTT GGC GGC ACT
M   S   *(Q) Y  L   K   R   R   K   K   L   A   Q   C   Q   R   F   G   G   T
61/21                                   91/31
AGT GGC ATT TCT TCA AGC CGC GAG ATT GTA GTG GGC TAT GTC GAT TCC CCT CCC ACG GAA
S   G   I   S   S   S   R   E   I   V   V   G   Y   V   D   S   P   P   T   E
121/41                                  151/51
GCT TTA AAA GAG TTA AAT AAA ACA GGC GAT ATT AAA ATA ATA AAA CAT TTA AAA GAA ATC
A   L   K   E   L   N   K   T   G   D   I   K   I   I   K   H   L   K   E   I
181/61                                  211/71
AAG GCA ATT GTA TTA AAC ATT CCC GAT AAT AAA ACA GAG AAA CTT AAG GAA AAG TTA AAA
K   A   I   V   L   N   I   P   D   N   K   T   E   K   L   K   E   K   L   K
241/81                                  271/91
GGA GTT AGA TAT ATA GAG GAA GAC GGC GTT GCG TAT GCG TTT GGT TTT TCT AAT TAT ACC
G   V   R   Y   I   E   E   D   G   V   A   Y   A   F   G   F   S   N   Y   T
301/101                                 331/111
GAT GTA CAG TGG AAT GTA AAA ATG ATA AAC GCC CCG CGT CTG GGA CGC CTA TTT TCT CAC
D   V   Q   W   N   V   K   M   I   N   A   P   R   L   G   R   L   F   S   H
361/121                                 391/131
ATT TGG CGA CGC GCA TTT GGC TAT GGA GTT AAA GTG GCG GTG CTC GAC ACA GGC ATT GAC
I   W   R   R   A   F   G   Y   G   V   K   V   A   V   L   D   T   G   I   D
421/141                                 451/151
TAC AAG CAC CCG GAG CTA TCC GGC AAG GTG GTT TAT TGT ATT AAC ACT CTC GGC AAC ACT
Y   K   H   P   E   L   S   G   K   V   V   Y   C   I   N   T   L   G   N   T
481/161                                 511/171
CTC TAC AAG GGG ACA AAT TTA AGG AAG TGC GCC GAC AGA AAA TGC CAC GGC ACG CAT GTA
L   Y   K   G   T   N   L   R   K   C   A   D   R   K   C   H   G   T   H   V
541/181                                 571/191
GCT GGG ATA ATA GCC GCT TCG TTG AAT AAC GTG AGC GCA GCC GGC GTT GTG CCT AAG GTG
A   G   I   I   A   A   S   L   N   N   V   S   A   A   G   V   V   P   K   V
601/201                                 631/211
CAG TTA ATA GCA GTT AAG GTC TTA TAC GAC AGC GGC TCC GGG TAC TAT AGC GAT ATT GCC
Q   L   I   A   V   K   V   L   Y   D   S   G   S   G   Y   Y   S   D   I   A
661/221                                 691/231
GAG GGG ATA ATA GAG GCA GTT AAA GCA GGG GCT TTA ATT CTA TCA ATG TCC CTA GGA GGC
E   G   I   I   E   A   V   K   A   G   A   L   I   L   S   M   S   L   G   G
721/241                                 751/251
CCC ACA GAC GCC TCT GTG TTG AGA GAC GCC TCG TAT TGG GCC TAT CAA CAA GGC GCT GTT
P   T   D   A   S   V   L   R   D   A   S   Y   W   A   Y   Q   Q   G   A   V
781/261                                 811/271
CAG ATA GCC GCC GCT GGT AAT TCA GGC GAT GGC GAT CCC TTG ACA AAC AAC GTG GGG TAT
Q   I   A   A   A   G   N   S   G   D   G   D   P   L   T   N   N   V   G   Y
841/281                                 871/291
CCC GCC AAG TAT AGC TGT GTA ATA GCA GCG GCG GCG GTA GAT CAA AAC GGC TCC GTC CCC
P   A   K   Y   S   C   V   I   A   A   A   A   V   D   Q   N   G   S   V   P
901/301                                 931/311
ACG TGG AGT AGC GAC GGG CCA GAG GTG GAC ACC GCG GCG CCA GGG GTA AAC ATA TTG TCC
T   W   S   S   D   G   P   E   V   D   T   A   A   P   G   V   N   I   L   S
961/321                                 991/331
ACA TAT CCC GGC GGC AGA TAC GCG TAT ATG TCC GGC ACA TCT ATG GCG ACG CCT CAC GTG
T   Y   P   G   G   R   Y   A   Y   M   S   G   T   S   M   A   T   P   H   V
1021/341                                1051/351
ACT GGC GTA GCG GCC TTA ATA CAA GCG TTG AGA CTC GCC TCA GGC AAG AGG TTG CTA ACC
T   G   V   A   A   L   I   Q   A   L   R   L   A   S   G   K   R   L   L   T
1081/361                                1111/371
CCA GAC GAG GTT TAT CAA GTA ATT ACC TCT ACG GCT AAG GAT ATC GGC CCG CCC GGT TTT
P   D   E   V   Y   Q   V   I   T   S   T   A   K   D   I   G   P   P   G   F
1141/381                                1171/391
GAC GTC TTT TCG GGC TAC GGC TTA GTT GAC GCA TAC GCC GCA GTT GTG GCC GCG CTA AGT
D   V   F   S   G   Y   G   L   V   D   A   Y   A   A   V   V   A   A   L   S
1201/401
CGC TAA CTTTTTATATAGAATTCAAATTGAGTATATGCCCACGTGGACTGAGTACATACTCTATAAAAAATTGGGGAA
R   *
AACTCCGTCGCCAGGTGACGTCGTTGAAATAGTTCCAGATCTCGTCGGCTTTCACGACTTGACGGGGTACCACGTCCTTG

AGGTGTTGGAAAGCATGGGCAAAGTGGAGGTGTTTGACAGGGAGAGAGTCGTTGTTGCGTTTGATCACTTGTCCCCGCCC

CCAAATCAGAGAGCCGCTGAGATAATGGTGTACATAAGGCGTCATGTCAAGGCTCTGGGCTTCCTAATTTCTACGACGTA

GGCGCGGCATTTTGCACCAGATTATCTGGAGAAATACGCCTTGCCGGGCCAAGTGATCTTCGCCGCGATAGCCACACTTT

CACCGCC
```

Fig. 2. Nucleotide sequence and corresponding amino acid sequence of the serine-type protease of *P. aerophilum*. The predicted leader and prepro sequences have been underlined.

The translated sequence showed a long ORF starting 83 amino acids upstream of sequence homology to various subtilisins. The first 15 amino acids encoded by this region showed similarities to leader sequences from subtilisins Carlsberg, BPN', 1168, and *T. aquaticus* aqualysin I (data not shown). The intervening region is not homologous to any known protein, and appears to be the N-terminal peptide autocatalytically cleaved on subtilisin's export from the cell (Terada et al., 1990). Based on these considerations, we assigned a potential initiator methionine for the protein, as well as a potential cleavage site for the mature protein (Fig. 2). The end of the gene is clearly defined by 5 stop codons. The first stop codon appeared 15 bases downstream, comparable in position to the end of the gene in several *Bacillus* species, and was followed by a poly T region.

Multiple sequence alignments

The similarity of the translated sequence to subtilisin encouraged us to perform a multiple sequence alignment using homologous proteins identified by BLAST searches. The programs CLUSTALV (Higgins & Sharp, 1989) and PredictProtein (Rost & Sander, 1993; send an e-mail message with "help" to PredictProtein@EMBL-Heidelberg.de) were used. The PredictProtein server generated a multiple sequence alignment and a secondary structure prediction for the *P. aerophilum* protein. The results are combined in Figure 3. To make the figure, the sequences were aligned with the PredictProtein output and known secondary structures of thermitase, Carlsberg, BPN', and proteinase K. The alignment similarity scores identified the *P. aerophilum* sequence as most similar to gram-positive subtilisins, but PredictProtein identified thermitase from *Thermoactinomyces vulgaris* (Meloun et al., 1985) as having the most similar structure. Similarity to other serine proteases was much weaker. In particular, the *P. aerophilum* sequence showed weak homology to aqualysin I, produced by *T. aquaticus* (Terada et al., 1990), and halolysin, a serine protease from a moderately thermophilic (60 °C) and halophilic archaeum (Kamekura et al., 1992). Neutral proteases such as thermolysin (Pauptit et al., 1988), despite their structural similarity, were not recovered by BLAST or PredictProtein and were not included in the alignment. Based on its observed similarities to other subtilisins, we have named the protein product of the *P. aerophilum* gene "aerolysin."

Potential thermophilic transitions in the Pyrobaculum subtilisin sequence

Recent studies (Frömmel & Sander, 1989; Menendez-Arias & Argos, 1989) have shown a pattern of substitutions associated with protein stability at high temperatures. In a preliminary analysis of the aerolysin sequence, sites potentially contributing to thermostability were scored manually using the statistically significant exchanges from Table 3 in Menendez-Arias and Argos (1989). A site was scored positive if the transition was predicted by the table and if it was shared by 3 or fewer of the 15 proteins in the alignment. Transitions were considered strongly thermophilic if the transitions were confined to the moderately thermophilic proteins thermitase, aqualysin I, halolysin, and proteinase K. A summary of the results is found in Table 1. The most common transition replaced a variety of amino acid types with alanine. This result fits several instances in their transition table, but runs counter to their finding that substitutions of increasing hydrophobicity are also stabilizing. In the *P. aerophilum* sequence, transitions to alanine were observed at at least 9 sites, with the only strong counterexamples found at Thr 258 and Val 260. By contrast, the Gly → Ala transition, reported as being statistically the most frequent by Argos et al. (1979), was found at only a single site in the protein. Suggestive evidence for the less statistically significant transitions Ser → Thr, Lys → Arg, and Asp → Glu were also found, although several counterexamples exist for these cases.

A second class of sites was also noted in which the *P. aerophilum* subtilisin replaced an uncharged amino acid with aspartic acid. The only subtilisins sharing these transitions were halolysin and TA41 from an Antarctic *Bacillus* species.

Model building

In order to further examine the structure of the transition sites in *P. aerophilum* aerolysin, 2 models were generated. The first, shown in Figure 4, maps data from Table 1 onto a cartoon of secondary structure drawn after Chen and Arnold (1993). The sequence/secondary structure alignments in Figure 3 were used to assign sites to the structure. Five of the alanine transitions map to the amino and carboxyl ends of helices H3, H4, and H7. Because these helices are on the protein surface, alanine replacement may result in more stable packing against the hydrophobic core and/or decreased helical flexibility. The counterexamples at positions 258–260 (where the *P. aerophilum* sequence has a bulky hydrophobic amino acid relative to other subtilisins) map to internal hydrophobic helix H5. Because this helix is buried, requirements for its stability may differ from those of surface helices.

The transitions to aspartic acid also show correlation with secondary structure, mapping to the N-terminal ends of helices H3, H4, H6, and H7 (Fig. 5C). Aspartic acid is common at the N-terminus of alpha-helices, where it can interact with the helix dipole to increase stability (Sali et al., 1988). However, the Antarctic subtilisin TA41 has even more negative charges than aerolysin in these positions, which raises doubts as to their role in thermostability.

The remaining sites map to variable surface loops and beta-strands E6 and E7. In order to interpret these sites, a simple 3-dimensional model was generated using the sequence alignment and Biosym's Homology program (see Materials and methods). Examination of the tertiary structure model reveals that several of the remaining sites may be in physical contact with each other (Fig. 5A,B). In thermitase, the turn containing Asp 47 passes near the loop containing Gly 88 and Ile 89. The corresponding residues in the *P. aerophilum* sequence all show thermophilic transitions, at Glu 62, Ser 109, and Ala 110, respectively. Although the true structure of the loop regions cannot be determined from the model, the clustering of thermostable sites may indicate that loops L4 and L6 continue to interact in *P. aerophilum* subtilisin. A similar spatial clustering is found for residues Ala 207, Ala 208, and Ala 229 (thermitase residues Val 181, Ala 182, and Ala 203). These residues are in the adjacent extended strands E6 and E7, connected by loop L8.

Localization and characterization of proteolytic activity

Analysis of the supernatant of a batch culture, of the cell envelope fraction, and the cytoplasmic fraction showed that about two-thirds of the proteolytic activity was associated with the cell envelope fraction and about one-third was found in the cytoplasm. No activity was present in the supernatant. Below 75 °C, no significant proteolytic activity was found. Activity was highest at neutral to alkaline pH and temperatures from 100 up to 130 °C (data not shown). Polyacrylamide gel substrate containing denaturing SDS showed a band at molecular weight 60 kDa after incubation for 30 min at 95 °C. After 60–90 min of incubation, up to 4 additional bands appeared. This is possibly due to self-digestion of the protein. SDS-PAGE therefore indicates that *P. aerophilum* contains at least 1 secreted protease with a size consistent with the aerolysin gene, which is active in the presence of SDS.

Discussion

Advantages of the study of hyperthermophilic proteins

Hyperthermophiles make interesting objects of study for several reasons. These organisms live under the most extreme conditions yet found for living organisms. Within the Archaea, some forms are active above 110 °C (Stetter et al., 1983; Kurr et al., 1991), which may be compared to the 50–70 °C temperature ranges for "classical" thermophiles such as *T. aquaticus*. In addition, it is generally believed that hyperthermophiles are ancestral to organisms that grow at room temperature. In the 16S rRNA phylogenetic tree, all known hyperthermophiles have short evolutionary lineages indicating slow evolution. The deepest branching organisms within 16S rRNA phylogenetic trees consist exclusively of hyperthermophilic bacteria and Archaea (Woese, 1990; Stetter, 1993). On this tree, the branch lengths of hyperthermophilic species are shorter than those of mesophiles. This indicates that these organisms evolved slowly relative to mesophiles and may be very similar to the progenote or the last common ancestor of all life.

Such considerations make proteins from hyperthermophiles of great interest for protein structure and engineering. If it is accepted that their thermal stability is a primitive and not an acquired feature, it is appropriate to study proteins from hyperthermophiles to determine mechanisms of protein stability. Structural studies may also be easier. Because the intramolecular forces in these proteins may be stronger than those in mesophiles, they may form crystals of greater compactness and stability. It is also worth noting that many of these proteins retain residual activity at mesophilic temperatures. Hyperthermophilic proteins are also good candidates for protein engineering. It seems likely that cold-active mutants of hyperthermophilic proteins could be created that retained some of their extreme stability at moderate temperatures. If this is the case, it may be easier to engineer low-temperature activity into hyperthermophilic proteins, instead of enhancing the thermostability of mesophilic proteins.

The P. aerophilum subtilisin gene

Analysis of the ORF encoding the protease gene revealed a number of interesting features. The N-terminal portion of the sequence encoded a likely leader peptide (Wiech et al., 1991) and a longer (~80 amino acids) precursor peptide. This suggests that the processing mechanism for the *P. aerophilum* protein is similar to that of other subtilisins. In these enzymes, a short leader peptide directs transport into the periplasmic space. A second, larger, "prepro" sequence allows folding into an active conformation and is subsequently cleaved by an autocatalytic process. Because proteolytic activity of *P. aerophilum* cells is confined to the cell envelope fraction, the *P. aerophilum* subtilisin is probably processed by a similar mechanism. It will be of interest to see whether active *P. aerophilum* subtilisin can be recovered when expressed in mesophiles such as *E. coli*. In other organisms, removal of the prepro peptide is independent of the primary sequence of the cleavage site (Egnell & Flock, 1992). It has been demonstrated that when aqualysin I from *T. aquaticus* is expressed in *Escherichia coli*, its N-terminal peptide is correctly removed and the precursor protein is exported to the periplasmic space (Terada et al., 1990). Subsequent heat treatment at 65 °C causes autocatalytic cleavage of the N-terminal peptide, activating the protease.

Secondary and tertiary structure analysis

Comparison of the sequence alignment and secondary structure prediction from PredictProtein indicates that *P. aerophilum* protease is structurally most similar to thermitase from *T. vulgaris*. However, 4 predicted surface loops (L1, L2, L3, and L10 in Figs. 3 and 4) are larger than their homologs in thermitase. L1 and L2 are comparable in size to the more distantly related proteins aqualysin I and proteinase K, while L3 and L10 are unique to aerolysin. Their presence is somewhat surprising, because large loops in the structure should increase the entropy of unfolding and destabilize the protein at high temperatures. A 3-dimensional structure will be needed to resolve the function of the expanded loops.

The sequence alignment also revealed individual sites within the protein that may contribute to its thermostability. The results are in contrast to a similar analysis of the *trpE*, *trpG*, and *trpD* genes from the hyperthermophilic eubacterium *Thermotoga maritima* (Kim et al., 1993), in which thermostability sites were difficult to identify. The most common change was alanine replacement, which appeared at at least 9 sites in *P. aerophilum* subtilisin (Table 1) and was especially common at the beginning and end of surface helices. Many of the same transitions were shared by thermitase, which functions at temperatures intermediate between the gram-positive subtilisins and aerolysin. The results are also consistent with the reported sequence of a subtilisin isolated from the Antarctic psychrotroph *Bacillus* TA41 (Davail et al., 1992). Compared to aerolysin, the corresponding sites in TA41 exhibit transitions likely to reduce stability.

In other positions, neutral amino acids at the N-terminus of helices H3, H4, H6, and H7 of mesophilic subtilisins were replaced by aspartic acid in aerolysin (Figs. 4, 5C). Three of these transitions were also found in halolysin, which is halophilic, moderately thermophilic, and the only other archaebacterial protease in the set. The local sequence similarity between halolysin and aerolysin in these regions contrasts with its lower overall homology to aerolysin. On the other hand, subtilisin TA41

```
                             10        20        30        40        50        60        70        80        90       100
AEROLYSIN          YIEEDGVAYAFGFSNYTD-VQWNVKMINAPASGTLFSHIWRRAF-GYGVKVAVLDTGIDYKHPELSGKVVYCINTLGNTLYKGTNLR-KCADRKCHGTHVAGIIAA-S
THERMITASE         ---------YTPNDPYFSSRQYGPQKIQAPQA-------WDIA-EGSGAKIAIVDTGVQSNHPDLAGKVVGG-----WDFVDNDSTP---QNGNGHGTHCAGIAAAVT
Halolysin          DVEYAEDNATYEAIATPNDPQYGQQYAPQQVNCEAA---WDVTYGDPGVTISVVDQGIQYDHEDLEGNMDGSVSNYGDDFVDNDGDPYPVSASENHGTHVGGIAAGGT
SUBT. TA41         TNKPEALYNAMA---ASQSTPWGIKAIYNNSN-------LTSTSGGAGINIAVLDTGVNTNHPDL--SNNVEQC--KDFTVGTNFTDNSCTDRQGHGTHVAGSALANG
SUBT. CARLSBERG    YVEEDHVAHALAL ---AQTVPYGIPLIKADKV-------QAQGFKGANVKVAVLDTGIQASHPDL--NVVGGAS-----FVAGEAYN---TDGNGHGTHVAGTVAAL-
SUBT. BACMS        YVEEDHIAHEY -----AQSVPYGISQIKAPAL-------HSQGYTGSNVKVAVIDSGIDSSHPDL--NVRGGAS-----FVPSETNPY--QDGSSHGTHVAGTIAAL-
SUBT. I168         YVEEDHIAHEY -----AQSVPYGISQIKAPAL-------HSQGYTGSNVKVAVIDSGIDSSHPDL--NVRGGAS-----FVPSETNPY--QDGSSHGTHVAGTIAAL-
SUBT. J            YVEEDHIAHEY -----AQSVPYGISQIKAPAL-------HSQGYTGSNVKVAVIDSG -IDSSHPDLNV RGGAS-----FVPSETNPY--QDGSSHGTHVAGTIAAL-
SUBT. DY           ---------------AQTVPYGIPLIKADKV-------QAQGYKGANVKVGIIDTGIAASHTDL--KVVGGAS-----FVSGESYN---TDGNGHGTHVAGTVAAL-
SUBT. BPN'         YVEEDHVAHAY -----AQSVPYGVSQIKAPAL-------HSQGYTGSNVKVAVIDSGIDSSHPDL--KVAGGAS-----MVPSETNPF--QDNNSHGTHVAGTVAAL-
ISP-I              -----------------DVNELPEGIKVIKA----PEMWAKGVKGKNIKVAVLDTGCDTSHPDLKNQIIGGKNFSDDDGGK-EDAI---SDYNGHGTHVAGTIAA-N
B. ALCALOPHILUS    YIEEDAEVTTM -----AQSVPWGISRVQAPAA-------HNRGLTGSGVKVAVLDTGI-STHPDL--NIRGGAS-----FVPGEPST---QDGNGHGTHVAGTIAAL-
ELASTASE YAB       YIEEDAEVTTM ------QTVPWGINRVQAPIA-------QSRGFTGTGVRVAVLDTGI-SNHADL--RIRGGAS-----FVPGEPNI---SDGNGHGTQVAGTIAAL-
BSUB MINOR PROT.FTAADSTDFKVLSDGTDTSDNFEQWNLEPIQVKQA----WKAGLTGKNIKIAVIDSGI-SPHDDL--SIAGGYS--------AVSYTSSYKDDNGHGTHVAGIIGAK-
Aqualysin I        FIEADKVVRAW ---ATQSPAPWGLDRIDQRDLPLSNSYTY--TATGRGVNVYVIDTGIRTTHREFGGRARVGYDALGG---------NGQDCNGHGTHVAGTIGGV-
PROTEINASE K       YIEQDAVVTIN ---AAQTNAPWGLARISSTSPGTSTYYY--DESAGQGSCVYVIDTGIEASHPEFEGRAQMV-----------KTYYYSSRDGNGHGTHCAGTVGSR-
PredictProtein     eee  hhhhhhE      LLL L              LLL   eee   L LL eEEEEEe LLLLLL   hhhhhhh       Lee   LLLLL LLLLLLLLL ee eeeee
Thermitase 2°                        THHHHTHHHHHHHTTHHHHT         TT     TT EEEEEE     TT TTT    TT         EETTTTE                HHHHHHHHHH
Proteinase K 2°                       EETT  HHHHHHT         EE     TTTTTT EEEEEE     TT HHHTT EEEEEE                               HHHHHHHHHH T
Carlsberg 2°                                THHHHHTHHHH         HHHT     TT EEEEEE     TT TT     EEEEEE        TT   TT             HHHHHHHHHH
BPN' 2°                                     HHHHHTTHHHHH         TT      TT EEEEEE     TT TT     EEEEEE        TT    TT            HHHHHHHHHH
```

```
                        110       120       130       140                150       160       170       180       190       200
AEROLYSIN          LNNVSAAGVVPKVQLIAVKVLYD-SGSGYYSDIAEGIIEAVKAGAL-------ILSMSLGGPRDASVLRDASYWAYQQGAVQIAAAGNSGDGDPLTNNVGYPAKYSCV
THERMITASE         NNSTGIAGTAPKASILAVRVL-DNSGSGTWTAVANGITYAADQGAK-------VISLSLGGTVGNSGLQQAVNYAWNKGSVVVAAAGNAG-----NTAPNYPAYYSNA
Halolysin          NNATGHAGIS-NCSLLSARAL-GDGGGGSLTDIADAIQWSADQGAD-------VINMSLGGGGFSQTLSNACEYAYNQGSLLVAAAGN-----GYGNSVSYPAAYDTV
SUBT. TA41         GTGSGVYGVAPEADLWAYKVL-GDDGSGYADDIAEAIRHAGDQATA--LNTKVVINMSLGSSGESSLITNAVDYAYDKGVLIIAAAGNSGPKPGSIGYPGALVNAVAV
SUBT. CARLSBERG    DNTTGVLGVAPSVSLYAVKVL-NSSGSGTYSGIVSGIEWATTNGMD-------VINMSLGGPSGSTAMKQAVDNAYARGVVVVAAAGNSGSSGN-TNTIGYPAKYDSV
SUBT. BACMS        NNSIGVLGVAPSSALYAVKVL-DSTGSGQYSWIINGIEWAISNNMD-------VINMSLGGPTGSTALKTVVDKAVSSGIVVAAAAGNEGSSGS-TSTVGYPAKYPST
SUBT. I168         NNSIGVLGVSPSASLYAVKVL-DSTGSGQYSWIINGIEWAISNNMD-------VINMSLGGPTGSTALKTVVDKAVSSGIVVAAAAGNEGSSGS-TSTVGYPAKYPST
SUBT. J            NNSIGVLGVSPSASLYAVKVL-DSTGSGQYSWIINGIEWAISNNMD-------VINMSLGGPSGSTALKTVVDKAVSSGIVVAAAAGNEGSSGS-SSTVGYPAKYPST
SUBT. DY           DNTTGVLGVAPNVSLYAIKVL-NSSGSGTYSAIVSGIEWATQNGLD-------VINMSLGGPSGSTALKQAVDKAYASGIVVVAAAGNSGSSGS-QNTIGYPAKYDSV
SUBT. BPN'         NNSIGVLGVAPSASLYAVKVL-GADGSGQYSWIINGIEWAIANNMD-------VINMSLGGPSGSAALKAAVDKAVASGVVVVAAAGNEGTSGS-SSTVGYPGKYPSV
ISP-I              DSNGGIAGVAPEASLLIVKVLGGENGSGQYEWIINGINYAVEQKVD-------IISMSLGGPSDVPELEEAVKNAVKNGVLVVCAAGNEGDGDERTEELSYPAAYNEV
B. ALCALOPHILUS    NNSIGVLGVAPNAELYAVKVL-GASGSGSVSSIAQGLEWAGNNGMH-------VANLSLGSPSPSATLEQAVNSATSRGVLVVAASGNSG-----AGSISYPARYANA
ELASTASE YAB       NNSIGVLGVAPNVDLYGVKVL-GASGSGSISGIAQGLQWAANNGMH-------IANMSLGSSAGSATMEQAVNQATASGVLVVAASGNSG-----AGNVGFPARYANA
BSUB MINOR PROT.HNGYGIDGIAPEAQIYAVKAL-DQNGSGDLQSLLQGIDWSIANRMD-------IVNMSLGTTSDSKILHDAVNKAYEQGVLLVAASGNDG----NGKPVNYPAAYSSV
Aqualysin I        --TYGVA---KAVNLYAVRVL-DCNGSGSTSGVIAGVDWVTRNHRR-----PAVANMSLGGGV-STALDNAVKNSIAAGVVYAVAAGNDNANACNYS----PARVAEA
PROTEINASE K       --TYGVA---KKTQLFGVKVL-DDNGSGQYSTIIAGMDFVASDKNNRNCPKGVVASLSLGGGY-SSSVNSAAARLQSSGVMVAVAAGNNNADARNYS----PASEPSV
PredictProtein      LLLeeee      eEEEEEEe    LLL    hHHHHhHHHHhhLLL          EEEEEeLLLL hHHHHHHHHHhh LL EEEEEe LLLLLLLLLL     L
Thermitase 2°                 TT EEEEEE     TT E    HHHHHHHHHHHHHTT              EEEE EEE      HHHHHHHHHHHHTT EEEEE                  ETT  TT
Proteinase K 2° T      T      TT EEEEEE     TT      HHHHHHHHHHHHHHHHH     TT     EEEEE  EE     HHHHHHHHHHHHTT EEEEE          EHHHEE    TTT TT
Carlsberg 2°                  TT EEEEEE     TT E    HHHHHHHHHHHHHTT              EEEE EEE      HHHHHHHHHHHHHT EEEEE           TT       ETTT TT
BPN' 2°                       TT EEEEEE     TT      HHHHHHHHHHHHHTT              EEEE   E      HHHHHHHHHHHHHT EEEEE           TT       ETTT TT
```

```
                        210       220       230       240                     250       260       270       280       290
AEROLYSIN          IAAAAVDQNGSVPTWSSDGPEVDTAAPGVNILSTYPGGR--------------YAYMSGTSMATPHVTGVAALIQALRLASGKRLLTPDEVYQVITSTAKDIGPPGFD
THERMITASE         IAVASTDQNDNKSSFSTYGSVVDVAAPGSWIYSTYPTST--------------YASLSGTSMATPHVAGVAGLLASQGRS-------ASNIRAAIENTADKISGTGT-
Halolysin          MAVSSLDEGETLSAFSNLGPEIELAAPGGNVLSSIPWDN--------------YDTFSGTSMASPVVAGVAGFTLSAHPN-----LSNAELRSHLQNTAVDVGLSSEE
SUBT. TA41         AALENTIQNGTYRVADFSSRGHKRTAGDYVIQKGDVEISAPGAAVYSTWFDGGYATISGTSMASPHAAGLAAKIWAQSPA-----ASNVDVRGELQTRASVN-DILSG
SUBT. CARLSBERG    IAVGAVDSNSNRASFSSVGAELEVMAPGAGVYSTYPTST--------------YATLNGTSMASPHVAGAAALILSKHPN-----LSASQVRNRLSSTATYLGSSF--
SUBT. BACMS        IAVGAVNSANQRASFSSAGSELDVMAPGVSIQSTLPGGT--------------YGAYNGTSMATPHVAGAAALILSKHPT-----WTNAQVRDRLESTATYLGSSF--
SUBT. I168         IAVGAVNSSNQRASFSSAGSELDVMAPGVSIQSTLPGGT--------------YGAYNGTSMATPHVAGAAALILSKHPT-----WTNAQVRDRLESTATYLGNSF--
SUBT. J            IAVGAVNSSNQRASFSSAGSELDVMAPGVSIQSTLPGGT--------------YGAYNGTSMATPHVAGAAALILSKHPT-----WTNAQVRDRLESTATYLGSSF--
SUBT. DY           IAVGAVDSNKNRASFSSVGAELEVMAPGVSVYSTYPSNT--------------YTSLNGTSMASPHVAGAAALILSKYPT-----LSASQVRNRLSSTATNLGDSF--
SUBT. BPN'         IAVGAVDSSNQRASFSSVGPELDVMAPGVSIQSTLPGNK--------------YGAYNGTSMASPHVAGAAALILSKHPN-----WTNTQVRSSLENTTTKLGDSF--
ISP-I              IAVGSVSVARELSEFSNANKEIDLVAPGENILSTLPNKK--------------YGKLTGTSMAAPHVSGALALIKSYEESFQRK--LSESEVFAQLIRRTLPLDIAK
B. ALCALOPHILUS    MAVGATDQNNNRASFSQYGAGLDIVAPGVNVQSTYPGST--------------YASLNGTSMATPHVAGAAALVKQKNPS-----WSNVQIRNHLKNTATSLGSTN--
ELASTASE YAB       MAVGATDQNNNRATFSQYGAGLDIVAPGVGVQSTVPGNG--------------YASFNGTSMATPHVAGVAALVKQKNPS-----WSNVQIRNHLKNTATNLGNTT--
BSUB MINOR PROT.VAVSATNEKNQLASFSTTGDEVEFSAPGTNITSTYLNQY--------------YATGSGTSQATPHAAAMFALLKQRDPA-----ETNVQLREEMRKNIVDLGTAGRD
Aqualysin I        LTVGATTSSDARASFSNYGSCVDLFAPGASIPSAWYTSDTAT------------QTLNGTSMATPHVAGVAALYLEQNP------SATPASVASAILNGATTGRLS--
PROTEINASE K       CTVGASDRYDRRSSFSNYGSVLDIFGPGTSILSTWIGGSTRS-------------ISGTSMATPHVAGLAAYLM-------TLGKTTAASACRYIADTANKGDLS-N
PredictProtein     EEEE   LLLL        LLL eeEEe L   eeEee                  eee        hHHHHHHHHHHHh  LLLLLL hHHHHHHHHh      LLLL
Thermitase 2°       EEEEEE TT  E TT     TT   EEEE    EEEEETTTE               EEEE EHHHHHHHHHHHHHH HHTTT           HHHHHHHHHT E   TTET
Proteinase K 2°     EEEEEE TT EE TT   E TT   EEEE    EEEEETTTEEEE                E  HHHHHHHHHHHHHHHH          TT    TTTHHHHHHT EE   E
Carlsberg 2°        EEEEEE TT  E TT     TT   EEEE    EEEEETTTE               EEEE  HHHHHHHHHHHHHHHH TT             HHHHHHHHHT E    HH
BPN' 2°             EEEEEE TT  E TT     TT   EEEE    EEEEETTTE               EEEE  HHHHHHHHHHHHHHHH T         T    HHHHHHHHHTT E   HH
```

```
                     300       310
AEROLYSIN          VFSGYGLVDAYAAVVAALSR*                  100
THERMITASE         -YWAKGRVNAYKAVQY*                       43
Halolysin          Q--GHGRVDAGQAVTTDPGDG****               32
SUBT. TA41         NSAGSGDDIASGFGFAKVQ*                    31
SUBT. CARLSBERG    -YYGKGLINVEAAAQ*                        46
SUBT. BACMS        -YYGKGLINVQAAAQ*                        44
SUBT. I168         -YYGKGLINVQAAAQ*                        43
SUBT. J            -YYGKGLINVQAAAQ*                        43
SUBT. DY           -YYGKGLINVEAAAQ*                        43
SUBT. BPN'         -YYGKGLINVQAAAQ*                        43
ISP-I              TLAGNGFLYLTAPDELAEKAEQSHLLTL*           38
B. ALCALOPHILUS    -LYGSGLVNAEAATR*                        43
ELASTASE YAB       -QFGSGLVNAEAATR*                        42
BSUB MINOR PROT.QQFGYGLIQYKAQATD*****                      34
Aqualysin I        -GIGSGSPNRLLYSLLSSGSG*                  31
PROTEINASE K       IPFGTVNLLAYNNYQA*                       33
PredictProtein             hhhHHHHHHHHh L
Thermitase 2°        TTE   E  HHHHHT
Proteinase K 2°      TT    EE
Carlsberg 2°         HHTT E   HHHHT
BPN' 2°              HHTT E   HHHHT
```

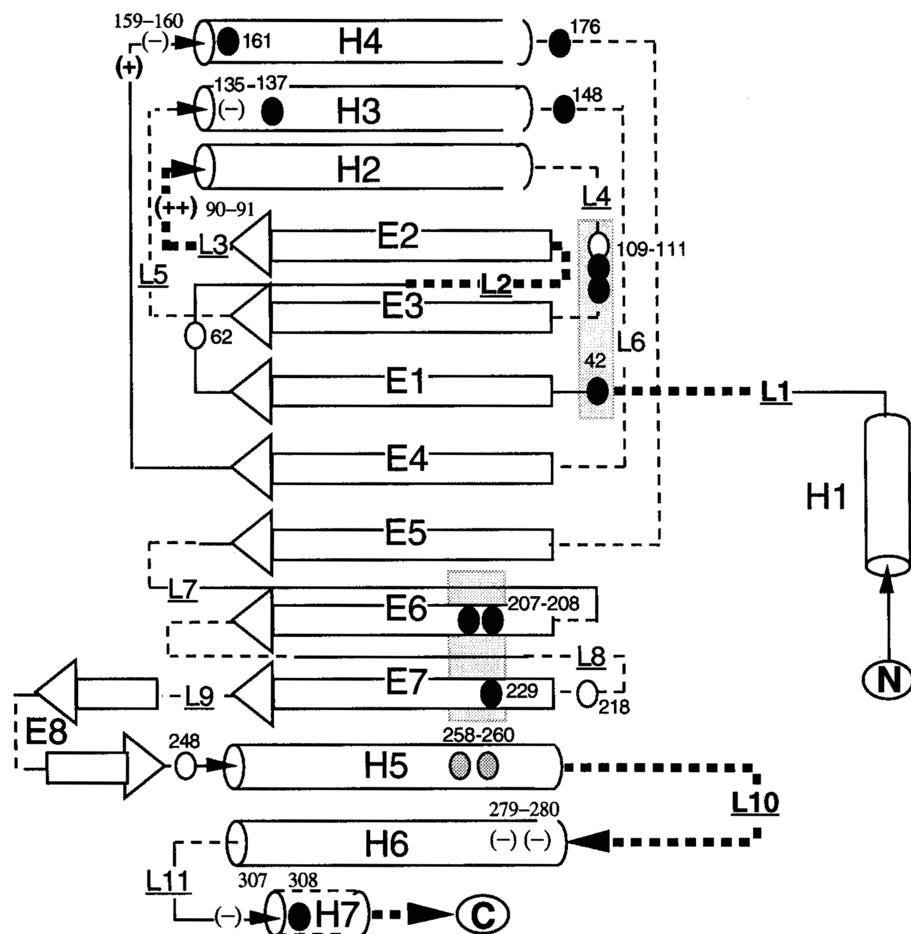


Fig. 4. Positions of sites potentially involved in thermostability in the *P. aerophilum* aerolysin. Structure cartoon is drawn after Chen and Arnold (1993); variable regions are from Siezen et al. (1993). Numbering and secondary structure designation ARE after the *P. aerophilum* subtilisin sequence in Figure 3. Heavy dashed lines indicate loop regions that are larger in the *P. aerophilum* sequence compared to Gram-positive subtilisins. Black circles indicate positions of thermostable transitions in which a bulky hydrophobic amino acid is replaced by alanine. Plus and minus signs indicate positions where the charge addition or replacement might affect the stability of an alpha-helix. Partly shaded circles indicate other transitions expected to have smaller effects on thermostability, after Menendez-Arias and Argos (1989). Clear circles indicate sites where the sequence transition is the reverse of that predicted by Menendez-Arias and Argos. Shaded regions indicate sites that may be in close proximity in the tertiary structure.

exhibited many of the same transitions. Instead of thermostability, the charge transitions may be related to the high ionic strength all 3 subtilisins encounter in their respective environments. Destabilizing charge transitions were observed N-terminal to helix H2, where *P. aerophilum* subtilisin has a small insertion (Asp-Arg-Lys) near the active site histidine. However, Arg and Lys have long side chains, and their conformations with respect to helix H2 are difficult to predict (Nicholson et al., 1988).

Fig. 3 (*facing page*). Sequence of the *P. aerophilum* protease (aerolysin), aligned with 14 subtilisins and subtilisin-like serine proteases. The first 16 lines show the alignment generated using the CLUSTAL algorithm. The bottom of the figure shows the secondary structure prediction from the PredictProtein algorithm, along with known secondary structures for several of the subtilisins in the alignment. Secondary structures are labeled with H = helix, E = extended, and L = loop/turn. Active site residues are boxed. Uppercase letters are "strong" (85% cutoff) predictions, lowercase letters are "weak" (75% cutoff) predictions according to the PredictProtein algorithm (for more information, send a help message to PredictProtein@EMBL-Heidelberg.de). Residues in bold represent the carboxyl portion of the prepro peptide cleaved from the mature protein, when known. Numbering of sequences used in the text and Table 1 proceeds from the first residue of the mature peptide when known. This residue is indicated in the alignment by double underlining. Otherwise, the first residue listed in the alignment is used for numbering. A **** at the sequence end indicates that a C-terminal portion of the protein is not included in the alignment. Pairwise similarity scores (percentile) are listed at the end of each sequence. AEROLYSIN, protease from *P. aerophilum*; THERMITASE, thermitase from *T. vulgaris* (Meloun et al., 1985); Halolysin, halolysin from an unnamed halophilic archaebacterium (Kamekura et al., 1992); SUBT. TA41, subtilisin from Antarctic *B. subtilis* strain TA41 (Davail et al., 1992); SUBT. CARLSBERG, subtilsin Carlsberg (Betzel et al., 1993); SUBT. BACMS, subtilisin Bacms (Svendsen et al., 1986); SUBT. I168, subtilisin I168 (Stahl & Ferrari, 1984); SUBT. J, subtilisin J (Jang et al., 1992); SUBT. DY, subtilisin DY (Betzel et al., 1993); SUBT. BPN′, subtilisin BPN′ (Pantoliano et al., 1989); ISP-I, ISP-1 (Rufo et al., 1990); B. ALCALOPHILUS, subtilisin from *Bacillus alcalophilus* (van der Laan et al., 1992); ELASTASE YAB, elastase YaB (Kaneko et al., 1989); BSUB MINOR PROT., minor protease from *B. subtilis* (Sloma et al., 1991); Aqualysin I, aqualysin I from *Thermus aquaticus* (Terada et al., 1990); PROTEINASE K, proteinase K from *Tritirachium album* (Gunkel & Gassen, 1989); PredictProtein, secondary structure prediction of PredictProtein algorithm for aerolysin; Thermitase 2°, Proteinase K 2°, Carlsberg 2°, BPN′ 2°, secondary structures of thermitase, proteinase K, subtilisin Carlsberg, and subtilisin BPN′.

Table 1. *Sites potentially important to the thermostability of P. aerophilum subtilisin*[a]

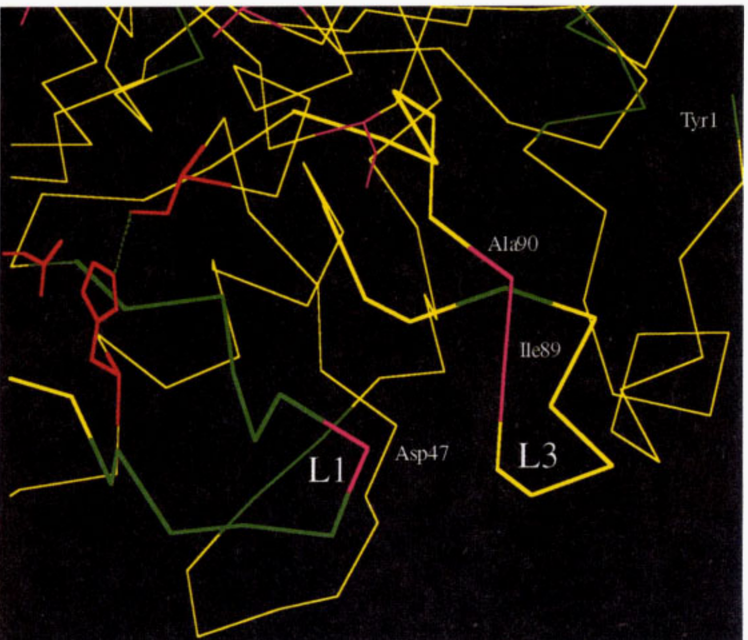
PA	TM	CB	TA41	Transition	Secondary structure
			Predicted from transition tables		
Ala 42	Ala 27	Gly 20	Thr 33	Hyd. → Ala	Turn C-terminal to helix H1
Glu 62	Asp 47	Asp 41	Asp 54	Asp → Glu	Turn in loop L2
Ser 109	Gly 88	Gly 79	Gly 99	Gly → Ser	Loop L4 (adj. to Glu 62)
Ala 110	Ile 89	Val 80	Val 100	Hyd. → Ala	Loop L4 (adj. to Glu 62)
Ala 111	Ala 90	Leu 81	Leu 101	Hyd. → Ala	Loop L4
Ala 137	Ala 116	Val 107	Ala 127	Hyd. → Ala	N-terminal helix H3
Ala 148	Ala 127	Met 118	Thr 138	Hyd. → Ala	C-terminal helix H3
Ala 161	Asn 140	Ser 131	Ser 156	Asn/Ser → Ala	Loop N-terminal to helix H4
Ala 176	Ser 155	Val 146	Val 171	Hyd. → Ala	Turn C-terminal to helix H4
Ala 207	Val 181	Val 176	Leu 202	Hyd. → Ala	Extended E6 (adj. to Ala 229)
Ala 208	Ala 182	Gly 177	Glu 203	Gly → Ala	Extended E6 (adj. to Ala 229)
Thr 218	Ser 192	Ser 187	Val 213	Ser → Thr	Turn in loop L8
Ala 229	Ala 203	Met 198	Thr 224	Hyd. → Ala	Extended E7 (adj. to Ala 181, 182)
Ser 248	Ser 220	Asn 217	Ser 257	Asn-Gly → Ser-Gly	Extended E8
Ala 308	Ala 273	Val 269	Ala 311	Hyd → Ala	N-terminal helix H7
			Counterpredictions		
Thr 258	Ala 232	Ala 227	Ala 260	Ala → Thr	Buried helix H5
Val 260	Val 234	Ala 229	Ala 262	Ala → Val	Buried helix H5
			Addition of negative charge at N-terminus of alpha-helices		
Asp 135	Thr 113	Gly 105	Asp 125	Unch. → (+)	N-terminal helix H3
Asp 160	Gly 139	Gly 130	Glu 155	Unch. → (+)	N-terminal helix H4
Asp 279	Ser 246	Ser 243	Val 283	Unch. → (+)	N-terminal helix H6
Glu 280	Asn 247	Gln 244	Asp 284	Unch. → (+)	N-terminal helix H6
Asp 307	Asn 272	Asn 268	Ile 310	Unch. → (+)	N-terminal helix H7
			Counterpredictions		
Arg 90	Gly 68	Gly 60	Arg 79	Unch. → (+)	N-terminal helix H2
Lys 91	Asn 69	Asn 61	Asn 80	Unch. → (+)	N-terminal helix H2

[a] Numbering is based on the first amino acid in the *P. aerophilum* aerolysin sequence alignment, rather than the position of the initiator methionine. The corresponding positions in thermitase (TM) from *T. vulgaris* (Meloun et al., 1985), subtilisin Carlsberg (CB; after Betzel et al., 1993), and subtilisin TA41 from the Antarctic *B. subtilis* strain TA41 (Davail et al., 1992) are given. Secondary structure descriptions are based on the Protein Data Bank 3-dimensional structure files for thermitase, subtilisins Carlsberg and BPN′, and proteinase K. Numbering proceeds from the first amino acid of the mature protein when known. Otherwise, the first residue listed in the alignment in Figure 3 is used. Helix numbering is after Figure 3. To create the table, the alignment shown in Figure 3 was used. Amino acid transitions were scored when they matched thermophilic transitions observed by Frömmel and Sander (1989) and Argos et al. (1979). For a site to be listed, (1) the *P. aerophilum* sequence contained the thermophilic transition, and (2) 3 or fewer of the nonthermophilic sequences carried the same transition. *P. aerophilum* subtilisin, halolysin, thermitase, and aqualysin I were classified as thermophilic.

In addition to these sites, other positions, mostly in variable loops, show transitions predicted by the statistical tables. Support for their role in protein stability comes from the observation that several of the proposed sites, while remote in the primary sequence, are adjacent in the tertiary structure model (Fig. 5A,B).

Calcium binding sites are also found in many other subtilisins (Frömmel & Sander, 1989; Gros et al., 1990; Teplyakov et al., 1990) and are believed to make a major contribution to stability. In order to determine whether aerolysin uses calcium binding to stabilize its structure, it will be necessary to characterize the protein at the physical level.

It is interesting to compare the *P. aerophilum* sequence with amino acid substitutions engineered into subtilisins to increase their stability. Pantoliano et al. (1989) made 6 concerted amino acid substitutions in subtilisin BPN′, causing nearly additive increases in thermal stability. The T_m for the protein was changed from 58.5 °C to 72.8 °C. In another study (Narhi et al., 1991), 3 mutations were made in the *aprA* gene of *Bacillus subtilis* to increase its long-term stability. Two of these mutations replaced asparagine residues with serines, preventing cyclization with adjacent glycine residues. A third substitution changed asparagine 76 in the high-affinity Ca^{2+} binding pocket to aspartic acid. These changes resulted in a significantly higher unfolding temperature and increased detergent resistance. Eijsink et al. (1992) increased the stability of a neutral protease by removing charged residues at the N-terminus of alpha-helices. With the exception of the Gly 169 → Ala change made by Pantoliano et al. (1989) in subtilisin BPN′, the *P. aerophilum* sequence does not conform to the replacements made in these studies. More recently, Chen and Arnold (1993) used sequential random mutagenesis to select a mutant subtilisin E able to function in 60% dimethylformamide. The 10 effective mutations clustered in variable loop regions on one side of the protein. Two of the mutations



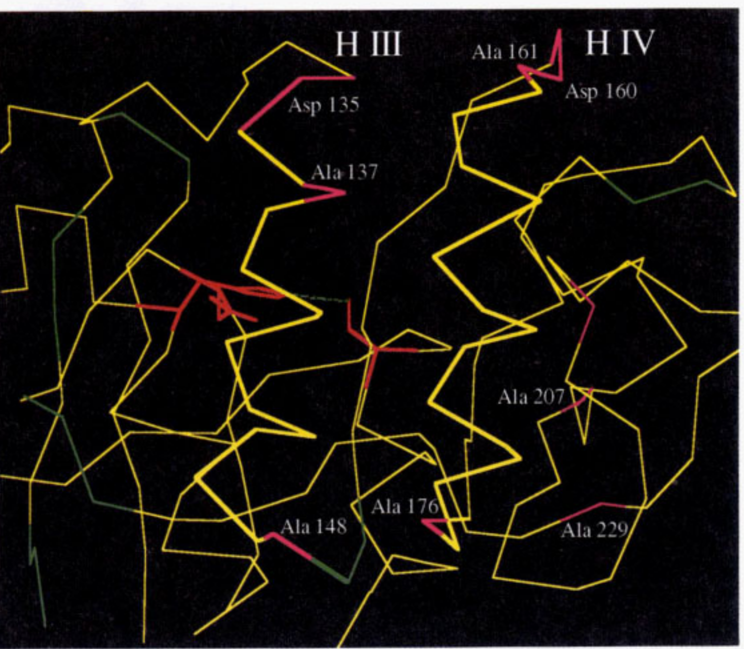


Fig. 5. Model of the potential 3-dimensional structure of the aerolysin protein, showing thermostable sites identified in the sequence alignment. The model was built using the Biosym Homology program with the tertiary structure of thermitase as a starting point. Residue numbering follows equivalent sites in the *P. aerophilum* sequence. Variable loop regions are shown in blue, and potential thermophilic sites are shown in red. Numbering of secondary structure elements is from Figure 3. **A:** Clustering of thermophilic residues from 2 surface loops, L1 and L3. **B:** Thermophilic sites in 2 adjacent extended strands, E6 and E7, linked by loop L8. **C:** Thermophilic sites, one each side of surface helices III and IV.

(Gly 160 → Asp in front of helix H4, and Asn 248 → Ser between E8 and H5) match a thermophilic transition in aerolysin. Most of the remaining changes were opposite the direction expected for thermostability, and are probably specific for activity in high concentrations of dimethylformamide.

The results of the sequence and structural analysis presented here make *P. aerophilum* aerolysin an attractive protein for expression, crystallization, and structure determination. Its similarity to other subtilisins, coupled with the extremely high temperatures at which it is active, will help our understanding of the origins of thermostability. The discovery of this protein in a relatively small-scale study of the *P. aerophilum* genome demonstrates the validity of the expressed sequence tag method for genome analysis of hyperthermophiles, and indicates that this organism could provide additional proteins valuable for research and biotechnology.

Materials and methods

Strains and media

P. aerophilum was grown aerobically in BS-medium at 97 °C as described previously (Völkl et al., 1993). Cells were harvested in the late exponential growth phase by centrifugation, and the cell masses were stored at −80 °C until use. *E. coli* strain XL1-Blue is described by the manufacturer (Stratagene, La Jolla, California) as *recA1*, *endA1*, *gyrA96*, *thi-1*, *supE44*, *relA1*, *lac* (*F'proAB lacI^q ZΔM15*, Tn*10*(*tet^r*)). *E. coli* strain SURE (Stratagene) is *mcrA*, Δ(*mcrBC-hsdRMS-mrr*)*171*, *supE44*, *thi-1*, *l*−, *gyrA96*, *relA*, *lac*, *recB*, *recJ*, *sbcC*, *umuC*::Tn*5* (*kan^r*), *uvrC* (*F'proAB lacI^{q}ZDM15*, Tn*10*(*tet^r*)).

Luria broth (LB) medium for growing the *E. coli* strains was prepared as described by Miller (1992). LB was supplemented with 0.2% maltose and 10 mM magnesium sulfate when strains were grown before and during phage infection. NCY broth contained, per liter of distilled water: 5 g NaCl, 2 g $MgSO_4 \cdot 7H_2O$, 5 g yeast extract (Difco), and 10 g NZ amine (casein hydrolysate; Sigma). The pH was adjusted to 7.5 with NaOH. Agar plates contained 1.5% agar (Difco), and top agarose contained 0.7% agarose (Bio-Rad) instead of agar. SM buffer contained 5.8 g NaCl, 2.0 g $MgSO_4 \cdot 7H_2O$, 50 mL 1 M Tris/HCl, pH 7.5, and 5 mL 2% gelatin per liter of water. Ampicillin was added to 100 μg/mL and tetracycline to 20 μg/mL medium as described by Miller (1992). Tetracycline was not added to media supplemented with maltose and magnesium sulfate.

Preparation of enzyme fractions

About 1 g of frozen cell paste was resuspended in 10 mL 50 mM Tris/HCl, pH 8.0, supplemented with 1 mM $CaCl_2$, and the cells were mechanically opened by vortexing in the presence of glass pads. The cell homogenate was separated into cytoplasmic and cell envelope fractions by centrifugation at 15,000 × *g* for 30 min in a Beckmann JA-20 rotor. The cell envelope fraction was washed twice in the same buffer as used above.

Protease assay

Proteolytic activity was determined by the hydrolysis of casein according to Kunitz (1947). If not mentioned otherwise, 100 μL of cell envelope fraction (CEF) was added to 0.5 mL of 0.6% casein (Merck) or 0.2% azocasein (Serva) in 50 mM Tris/HCl, pH 8.0. After incubation, the reaction was stopped by adding 1 mL of 5% trichloroacetic acid and incubating at room temperature for about 30 min. The reaction was centrifuged in an Eppendorf centrifuge for 10 min, and the absorbance of the supernatant was determined at 280 nm when casein was used and 420 nm when azocasein was used.

Substrate containing SDS

Discontinuous SDS-PAGE was performed according to Laemmli (1970), except that 0.01% casein (Merck) and 0.1% SDS were added to the 10% polyacrylamide separation gel. The 3% stacking gel was supplemented with 0.1% SDS only. Electrophoresis, incubation, and fixing of the gel were done as described by Connaris et al. (1991) except that the gel was incubated at 95 °C in 50 mM Tris/HCl, 1 mM $CaCl_2$ buffer, pH 8.0.

DNA isolation and characterization

DNA from selected λ clones was isolated according to Sambrook et al. (1989). Plasmid DNA was isolated by the alkaline lysis procedure as described by Kraft et al. (1988). Prior to sequencing, the DNA was analyzed by restriction mapping and agarose gel analysis as described by Sambrook et al. (1989).

Plaque lift, colony lift, and Southern hybridizations of agarose gels were carried out according to Sambrook et al. (1989) using membranes from Bio-Rad. For the hybridizations, 18-mer oligonucleotide probes radiolabeled with γ^{32}P-dATP were used. Hybridization was performed by shaking overnight at 50 °C. The membranes were washed in 20 mM Na_2HPO_4, 5% SDS at 50 °C for 45 min, covered with Saran Wrap, and exposed to Kodak XAR film.

Sequencing and data analysis

Clones from the oligo-dT primed library and genomic libraries were sequenced by the Sanger chain termination method using a Sequenase version 2.0 kit (US Biochemical) and α^{32}P-dATP (New England Nuclear). Sequencing products were separated on 6% polyacrylamide–urea gels at 2 intervals to obtain overlapping sequencing runs.

Sequences were analyzed for similarity to known proteins with the NCBI BLAST program (Altschul et al., 1990) using the NCBI nonredundant database containing the GenBank, PIR, Swiss-Prot, and EMBL databases. Database sequences were retrieved using BLAST retrieve or Internet Gopher (Krol, 1992). Multiple sequence alignments were performed using CLUSTALV (Higgins & Sharp, 1989), and secondary structure predictions based on alignments used the PredictProtein server (Rost & Sander, 1993). Using these 2 sources, a final alignment was derived. Three-dimensional models of the *P. aerophilum* subtilisin were created using the InsightII and Homology programs from Biosym Technologies (San Diego, California). The structurally conserved regions of thermitase were used for creating the model, but refinement of the loop structures and other structures was not attempted. In a previous study comparing predicted and ac-

tual structures for thermitase, it was found that molecular dynamics and energy minimization were inadequate to select the correct loop conformation for subtilisins (Frömmel & Sander, 1989).

Acknowledgments

We thank Jean J. Lee for excellent technical assistance. This work was supported in part by grants from the Office of Naval Research (grant N00014-92-5-1403) to J.H.M. and from the Deutsche Forschungsgemeinschaft to K.O.S.

References

Adams MD, Soares MB, Kerlavage AR, Fields C, Venter JC. 1993. Rapid cDNA sequencing from a directionally cloned human infant brain cDNA library. *Nature Genet 4*:373–380.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215*:403–410.

Argos P, Rossmann MG, Grau UM, Zuber H, Frank G, Tratschin JD. 1979. Thermal stability of proteins. *Biochemistry 18*:5698–5703.

Betzel C, Visanji M, Eschenburg S, Wilson KS, Peters K, Fittkau S, Singh TP, Genov N. 1993. Crystallization and preliminary X-ray analysis of subtilisin DY, a natural mutant of subtilisin Carlsberg. *Arch Biochem Biophys 302*:499–502.

Burggraf S, Larsen N, Woese CR, Stetter KO. 1993. An intron within the 16S ribosomal RNA gene of the archaeum *Pyrobaculum aerophilum*. *Proc Natl Acad Sci USA 90*:2547–2550.

Chen K, Arnold FH. 1993. Tuning the activity of an enzyme for unusual environments: Sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proc Natl Acad Sci USA 90*:5618–5622.

Connaris H, Cowan DA, Sharp RJ. 1991. Heterogeneity of proteinases from the hyperthermophilic archaeobacterium *Pyrococcus furiosus*. *J Gen Microbiol 137*:1193–1199.

Davail S, Feller G, Narinx E, Gerday C. 1992. Sequence of the subtilisin-encoding gene from an Antarctic psychrotroph *Bacillus* TA41. *Gene 119*:143–144.

Egnell P, Flock JI. 1992. The autocatalytic processing of the subtilisin Carlsberg pro-region is independent of the primary structure of the cleavage site. *Mol Microbiol 6*:1115–1119.

Eijsink VGH, Vriend G, van den Burg B, van der Zee JR, Venema G. 1992. Increasing the thermostability of a neutral protease by replacing positively charged amino acids in the N-terminal turn of alpha-helices. *Protein Eng 5*:165–170.

Frömmel C, Sander C. 1989. Thermitase, a thermostable subtilisin: Comparison of predicted and experimental structures and the molecular cause of thermostability. *Proteins Struct Funct Genet 5*:22–37.

Gros P, Kalk KH, Hol WGJ. 1990. Calcium binding to thermitase: Crystallographic studies of thermitase at 0, 5, and 100 mM calcium. *J Biol Chem 266*:2953–2961.

Gunkel FA, Gassen HG. 1989. Proteinase K from *Tritirachium album* Limber. Characterization of the chromosomal gene and expression of the cDNA in *Escherichia coli*. *Eur J Biochem 179*:185–194.

Higgins DG, Sharp PA. 1989. Fast and sensitive multiple sequence alignments on a microcomputer. *Comput Appl Biosci (CABIOS) 5*:151–153.

Kamekura M, Seno Y, Holmes ML, Dyall-Smith ML. 1992. Molecular cloning and sequencing of the gene for a halophilic alkaline serine protease (halolysin) from an unidentified halophilic Archaea strain (172P1), and expression of the gene in *Haloferax volcanii*. *J Bacteriol 174*:736–742.

Kaneko R, Koyama N, Tsai YC, Juang RY, Yoda K, Yamasaki M. 1989. Molecular cloning of the structural gene for alkaline elastase YaB, a new subtilisin produced by an alkalophilic *Bacillus* strain. *J Bacteriol 171*:5232–5236.

Kim CW, Markiewicz PM, Lee JJ, Schierle CF, Miller JH. 1993. Studies of the hyperthermophile *Thermotoga maritima* by random sequencing of cDNA and genomic libraries: Identification and sequencing of the *trp EG(D)* operon. *J Mol Biol 231*:960–981.

Kraft R, Tardiff J, Krauter KS, Lainwand LA. 1988. Using mini-prep plasmid DNA for sequencing double-stranded templates with sequenase. *BioTechniques 6*:544–546.

Krol E. 1992. Tunneling through the Internet: Gopher. In: *The whole Internet: User's guide and catalog*. Sebastopol, California: O'Reilly & Associates, Inc. pp 189–210.

Kunitz M. 1947. Crystalline soybean trypsin inhibitor II. General properties. *J Gen Physiol 30*:291–310.

Kurr M, Huber R, König H, Jannasch HW, Fricke H, Trincone A, Kristjansson JK, Stetter KO. 1991. *Methanopyrus kandleri*, gen. and sp. nov. represents a novel group of hyperthermophilic methanogens, growing at 110 °C. *Arch Microbiol 156*:239–247.

Laemmli UK. 1970. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature (Lond) 227*:680–685.

Meloun B, Baudys M, Kostka V, Hausdorf G, Frömmel C, Höhne WE. 1985. Complete primary structure of thermitase from *Thermoactinomyces vulgaris*, and its structural features related to the subtilisin-type proteinases. *FEBS Lett 183*:195–200.

Menendez-Arias L, Argos P. 1989. Engineering protein thermal stability. Sequence statistics point to residue substitutions in alpha-helices. *J Mol Biol 206*:397–406.

Miller JH. 1992. *A short course in bacterial genetics*. Cold Spring Harbor, New York: Cold Spring Harbor Press. pp 17–43.

Mitchinson C, Wells JA. 1989. Protein engineering of disulfide bonds in subtilisin BPN'. *Biochemistry 28*:4807–4815.

Narhi LO, Stabinsky Y, Levitt M, Miller L, Sachdev R, Finley S, Park S, Kolvenbach C, Arakawa T, Zukowski M. 1991. Enhanced stability of subtilisin by three point mutations. *Biotechnol Appl Biochem 12*:12–24.

Nicholson H, Bektel WJ, Matthews BW. 1988. Enhanced protein thermostability from designed mutations that interact with alpha-helix dipoles. *Nature (Lond) 336*:651–656.

Pantoliano MW, Whitlow M, Wood JF, Dodd SW, Hardman KD, Rollence ML, Bryan PN. 1989. Large increases in general stability for subtilisin BPN' through incremental changes in the free energy of unfolding. *Biochemistry 28*:7205–7213.

Pauptit R, Karlsson R, Picot D, Jenkins AS, Niklaus-Reimer J, Jansonius K. 1988. Crystal structure of neutral protease from *Bacillus cereus* refined at 3.0 Å resolution and comparison with the homologous but more thermostable enzyme thermolysin. *J Mol Biol 199*:525–537.

Rost B, Sander C. 1993. Prediction of protein secondary structure at better than 70%. *J Mol Biol 232*:584–599.

Rufo GA Jr, Sullivan BJ, Sloma A, Pero J. 1990. Isolation and characterization of a novel extracellular metalloprotease from *Bacillus subtilis*. *J Bacteriol 172*:1019–1023.

Sali D, Bycroft M, Fersht AR. 1988. Stabilization of protein structure by interaction of alpha-helix dipole with a charged side chain. *Nature 335*:740–743.

Sambrook J, Fritsch E, Maniatis T. 1989. *Molecular cloning: A laboratory manual*. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.

Schuliger JW, Brown SH, Baross JA, Kelly RM. 1993. Purification and characterization of a novel amylolytic enzyme from ES4, a marine hyperthermophilic archaeum. *Mol Mar Biol Biotechnol*. Forthcoming.

Siezen RJ, Bruinenberg PG, Vos P, van Alen-Boerrigter, Nijhuis M, Alting AC, Exterkate FA, de Vos WM. 1993. Engineering of the substrate-binding region of the subtilisin-like, cell-envelope proteinase of *Lactococcus lactis*. *Protein Eng 6*:927:937.

Sloma A, Rufo GA Jr, Theriault KA, Dwyer M, Wilson SW, Pero J. 1991. Cloning and characterization of the gene for an additional extracellular serine protease of *Bacillus subtilis*. *J Bacteriol 173*:6889–6895.

Stahl ML, Ferrari E. 1984. Replacement of the *Bacillus subtilis* subtilisin structural gene with an in vitro derived deletion mutation. *J Bacteriol 158*:411–418.

Stetter KO. 1993. Life at the upper temperature border. In: Tran Than Van J, Tran Than Van K, Mounolou JC, Schneider J, McKay C, eds. *Frontiers of life. Colloque Interdisciplinaire du Comité National de la Recherche Scientifique, C55*. Gif-sur-Yvette, France: Editions Fronières. pp 195–219.

Stetter KO, König H, Stackebrandt E. 1983. *Pyrodictium* gen. nov, a new genus of submarine disc-shaped sulfur reducing archaebacteria growing optimally at 105 °C. *Syst Appl Microbiol 4*:535–551.

Svendsen I, Genov N, Idakieva K. 1986. Complete amino acid sequence of alkaline mesenteric peptidase: A subtilisin isolated from a strain of *Bacillus mesentericus*. *FEBS Lett 196*:228–232.

Teplyakov AV, Kuranova IP, Harutyunyan EH, Vainshtein BK. 1990. Crystal structure of thermitase at 1.4 Å resolution. *J Mol Biol 214*:261–279.

Terada I, Kwon ST, Miyata Y, Matsuzawa H, Ohata T. 1990. Unique precursor structure of an extracellular protease, aqualysin I, with NH_2 and COOH-terminal Pro-sequences and its processing in *Escherichia coli*. *J Biol Chem 265*:6576–6581.

van der Laan JM, Teplyakov AV, Kelders H, Kalk KH, Misset O, Mulleners LJ, Dijkstra BW. 1992. Crystal structure of the high-alkaline serine protease PB92 from *Bacillus alcalophilus*. *Protein Eng 5*:405–411.

Völkl P, Huber R, Drobner E, Rachel R, Burggraf S, Trincone A, Stetter

KO. 1993. *Pyrobaculum aerophilum* sp. nov., a novel nitrate-reducing hyperthermophilic archaeum. *Appl Environ Microbiol 59*:2918–2926.
Wells JA, Powers DB, Bott RR, Katz BA, Ultsch MH, Kossiakoff AA, Power SD, Adams RM, Heyneker HH, Cunningham BC, Miller JV, Graycar TP, Estell DA. 1987. Protein engineering of subtilisin. In: Oxender DL, Fox CF, eds. *Protein engineering*. New York: Alan R. Liss. pp 279–287.
Wiech H, Klappa P, Zimmerman R. 1991. Protein export in prokaryotes and eukaryotes. Theme with variations. *FEBS Lett 285*:182–188.
Woese CR. 1990. Evolutionary questions: The "progenote." *Science 247*:789.
Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA 87*:4576–4579.

Forthcoming Papers

Refined structure of dimeric diphtheria toxin at 2.0 Å resolution
M.J. Bennett, S. Choe, and D. Eisenberg

Refined structure of monomeric diphtheria toxin at 2.3 Å resolution
M.J. Bennett and D. Eisenberg

Pseudodihedrals: Simplified protein backbone representation with knowledge-based energy
R.S. DeWitte and D.E.I. Shakhnovich

Identification of a molecular switch that selects between two crystal forms of bovine pancreatic trypsin inhibitor
W.H. Gallagher and K.M. Croker

The chaperonin from the archaeon *Sulfolobus solfataricus* promotes correct refolding and prevents thermal denaturation in vitro
A. Guagliardi, L. Cerchia, S. Bartolucci, and M. Rossi

Anatomy of an engineered NAD-binding site
P.R.E. Mittl, A. Berry, N.S. Scrutton, R.N. Perham, and G.E. Schulz

The NMR solution structure of the pheromone E*r*-2 from the ciliated protozoan *Euplotes raikovi*
M. Ottiger, T. Szyperski, P. Luginbühl, C. Ortenzi, P. Luporini, R.A. Bradshaw, and K. Wüthrich

Structure and functions of arrestins
K. Palczewski

Optimization of the electrostatic interactions in proteins of different functional and folding type
V.Z. Spassov, A.D. Karshikoff, and R. Ladenstein

Dimerization of βB2-crystallin: The role of the linker peptide and the N- and C-terminal extensions
S. Trinkl, R. Glockshuber, and R. Jaenicke

Interactions between human defensins and lipid bilayers: Evidence for formation of multimeric pores
W.C. Wimley, M.E. Selsted, and S.H. White