Serpins in Prokaryotes

James A. Irving,† Peter J. M. Steenbakkers,‡ Arthur M. Lesk,*§ Huub J. M. Op den Camp,‡ Robert N. Pike,* and James C. Whisstock*†*

*Department of Biochemistry and Molecular Biology, Monash University, Melbourne; †Victorian Bioinformatics Consortium, Monash University, Melbourne; ‡Department of Microbiology, University of Nijmegen, The Netherlands; and §Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge Clinical School, U.K.

Members of the serpin (serine proteinase inhibitor) superfamily have been identified in higher multicellular eukaryotes (plants and animals) and viruses but not in bacteria, archaea, or fungi. Thus, the ancestral serpin and the origin of the serpin inhibitory mechanism remain obscure. In this study we characterize 12 serpin-like sequences in the genomes of prokaryotic organisms, extending this protein family to all major branches of life. Notably, these organisms live in dramatically different environments and some are evolutionarily distantly related. A sequence-based analysis suggests that all 12 serpins are inhibitory. Despite considerable sequence divergence between the proteins, in four of the 12 sequences the region of the serpin that determines proteinase specificity is highly conserved, indicating that these inhibitors are likely to share a common target. Inhibitory serpins are typically prone to polymerization upon heating; thus, the existence of serpins in the moderate thermophilic bacterium *Thermobifida fusca*, the thermophilic bacterium *Thermoanaerobacter tengcongensis*, and the hyperthermophilic archaeon *Pyrobaculum aerophilum* is of particular interest. Using molecular modeling, we predict the means by which heat stability in the latter protein may be achieved without compromising inhibitory activity.

Introduction

Serpins are members of a large family of proteinase inhibitors (about 500 identified to date; Irving et al. 2000) that inactivate target proteinases using a unique ''inhibition by distortion'' mechanism (Huntington, Read, and Carrell 2000). Whereas the majority of serpins target serine proteinases, several members have been discovered that are capable of inhibiting other classes of proteinases (e.g., the viral serpin crmA with caspase-1 [Ray et al. 1992] and SCCA-1 and MENT with papain-like cysteine proteinases [Schick et al. 1998; Irving et al. 2002]). In addition, several serpins have been identified that do not perform an inhibitory role—for example, the 47-kDa heat shock protein acts as a molecular chaperone (Nakai et al. 1992). Serpins have been characterized in many cellular and extracellular contexts, participating in processes such as blood clotting, intracellular signaling, prevention of tissue injury, and hormone binding (Silverman et al. 2001).

In contrast to small ''rigid'' proteinase inhibitors such as those of the Kazal or Kunitz family, inhibitory serpins rely on a complex conformational change to inhibit the target proteinase (Huntington, Read, and Carrell 2000). This mechanism requires precise timing to capture the target proteinase: the serpin fold has evolved to function as a finely tuned molecular machine. Some mutations that alter the conformational mobility of the serpin scaffold can result in the spontaneous transition to inactive forms, such as polymers or the latent conformation (see Stein and Carrell 1995). Other mutations interfere with the process of conformational change and impair serpin function by promoting substrate-like rather than inhibitory behavior (Hopkins, Carrell, and Stone 1993; Stein and Carrell 1995).

A fundamental, unanswered question in the field is the evolutionary origin of the serpin fold, which is closely related to the origin of the serpin mechanism itself. In a previous study, we performed an extensive phylogenetic investigation of the serpin superfamily (Irving et al. 2000). Despite using sensitive database searching methods such as PSI-BLAST and Hidden Markov Models, we were able to identify serpins only in viruses and higher (multicellular) eukaryotes and were unable to identify any putative prokaryotic or fungal serpin. This observation was somewhat surprising because the presence of serpins in both the animal and plant kingdoms suggests that they should also be found in a common ancestor (e.g., fungi or prokaryotes). Information bearing on the question of whether the ancestral serpin was a proteinase inhibitor would provide insight into the origin of the inhibitory mechanism: did this mechanism develop with the serpin fold or did the first serpin fulfill a noninhibitory function? Current biochemical evidence does not provide a simple answer. The most common serpin targets, trypsin-like serine proteinases, have bacterial homologs (HtrA; Lipinska, Zylicz, and Georgopoulos 1990), and serpins have shown activity against bacterial subtilisin (Komiyama et al. 1996; Dahlen, Foster, and Kisiel 1997) and the caspase-like gingipain K enzyme from *Porphyromonas gingivalis* (Snipas et al. 2001). These enzymes are present in organisms that lack recognizable serpin sequences.

In this study we have identified, from recently released genomic data, the first examples of prokaryotic serpins. Two closely related cyanobacteria, *Nostoc punctiforme* and *Anabaena* sp. PCC 7120, the firmicutes

Abbreviations: P_n notation, according to the convention of Schecter and Berger (1967), residues of a peptide substrate N-terminal to the scissile bond are denoted P_n, . . . , P_2, P_1 and those C-terminal to the scissile bond are denoted P_1', P_2', . . . , P_m' RCL, reactive center loop.

Key words: serpin, prokaryote, comparative genomics, proteinase inhibitor, phylogeny, proteinase.

Address for correspondence and reprints: James C. Whisstock, Department of Biochemistry and Molecular Biology, Monash University, Clayton Campus, Melbourne, Victoria 3800, Australia. E-mail: james.whisstock@med.monash.edu.au.

Mol. Biol. Evol. 19(11):1881–1890. 2002

Thermobifida fusca, *Desulfitobacterium hafniense*, *Ruminococcus albus*, and *Thermoanaerobacter tengcongensis*, the green nonsulfur bacteria *Dehalococcoides ethenogenes*, the crenarchaeote *Pyrobaculum aerophilum*, and the euryarchaeotes *Methanosarcina acetivorans* and *Methanosarcina mazei* all contain putative serpin sequences. Sequence motifs and patterns of conservation suggest that they adopt a serpin-like fold and are capable of proteinase inhibition. The distribution of prokaryotic serpin sequences is sporadic. Although the serpin-bearing prokaryotes occur in markedly different environments, it remains unclear as to whether these genes have arisen through lateral gene transfer from eukaryotes or through inheritance of an ancient, prokaryotic serpin ancestor. The existence of a serpin in a hyperthermophilic organism is of particular interest. Using molecular modeling with reference to experimental evidence, we suggest a means by which resistance to polymerization can be obtained without compromising inhibitory activity.

Materials and Methods

Database Searching

A BLAST search was performed, using the protein sequence of human α_1-antitrypsin and the tblastn program, against the finished and unfinished microbial genomes at the NCBI (http://www.ncbi.nlm.nih.gov/blast). Seven annotated nucleotide sequences from *Anabaena* sp. PCC 7120 (Kaneko et al. 2001), *P. aerophilum* (Fitz-Gibbon et al. 2002), *M. acetivorans* (Galagan et al. 2002), *M. mazei* (Deppenmeier et al. 2002), *T. tengcongensis* (Bao et al. 2002), and five unannotated draft sequences were identified with an expect score less than 1×10^{-3}. The preliminary genome data were obtained from the DOE Joint Genome Institute (JGI; http://www.jgi.doe.gov) and The Institute for Genomic Research (TIGR; http://www.tigr.org). Nucleotide sequence of high quality was obtained for *N. punctiforme* and *T. fusca* serpins (phrap score of 90, or 10^{-7} errors per 100 base pairs) and of reasonable quality for *D. hafniense* (phrap score of 37–90). Information on sequence quality was not available for *D. ethenogenes* or *R. albus*. For each putative serpin, the amino acid translation was used to probe the database a second time but failed to identify further homologous sequences.

Sequence Analysis and Phylogenetic Relationships

The putative serpins were aligned and phylogenetic trees constructed as described previously (Irving et al. 2000). Briefly, a structure-based alignment was generated of human α_1-antitrypsin (pdb accession 1qlp), human antithrombin (1azx), *Manduca sexta* serpin 1K (1sek), and chicken ovalbumin (1ova). The bacterial serpins were aligned to this ''seed'' alignment with reference to secondary structure-specific gap penalties using CLUSTALW (Higgins, Thompson, and Gibson 1996). Sites in the alignment were compared with sites strictly conserved in >70% of the serpin superfamily (identified in Irving et al. 2000).

A profile alignment was used to incorporate these sequences into an alignment of 219 sequences from the superfamily as a whole, which in turn was used to derive bootstrapped neighbor-joining distance-based trees (500 replicates, JTT substitution model, using MOLPHY; Adachi and Hasegawa 1996). Clusters of evolutionarily related sequences were obtained after analysis of the bootstrapped trees using the majority consensus tree approach (Felsenstein 1985), the comparison consensus method, and the tree division method (Irving et al. 2000). For the tree division method, the ''clade-defining'' proteins were from (1) vertebrates: human α_1-antitrypsin (GenPept accession P01009), heat shock protein 47 (P29043), neuroserpin (Q99574), heparin cofactor II (AAC16324), alpha-2-antiplasmin (Q95121), angiotensinogen (P11859), and ovalbumin (P01012); (2) arthropods: Limulus intracellular coagulation inhibitor-1 (BAA06909) and *M. sexta* serpin 1K (AAB58491); (3) nematodes: *Caenorhabditis elegans* (AAB37049); (4) plants: barley protein Z (CAA66232), *Triticum aestivum* (CAA72274) and *Arabidopsis thaliana* (AAC27146).

Local Sequence Clusters

Predicted genes at genome positions 892,000–912,000 (*Anabaena* sp. PCC 7120), 15,000–35,000 (*P. aerophilum*), 2,750,757–2,773,734; 3,235,948–3,255,573; and 4,189,820–4,190,826 (*M. acetivorans*), 1,508,868–1,529,055 (*T. tengcongensis*), and sections 295–297 (accessions AE013513–AE013515, *M. mazei*) were obtained from the GenBank database. Nucleotide sequence from the region surrounding each putative serpin was obtained for *T. fusca* (within 10 kb upstream and downstream, on contig 64), *N. punctiforme* (contig 423), and *D. hafniense* (contig 3204). Each region was then compared with the nonredundant database using the standalone blastx program (genetic code 11; maximum expect 1×10^{-3}; otherwise default parameters) and potential genes were thereby inferred.

Search for Proteinases Correlating with the Presence of Serpins

Sequence alignments representative of all proteinase families in the MEROPS database (http://www.merops.co.uk; Rawlings, O'Brien, and Barrett 2002) were obtained. A consensus sequence was derived from each alignment, and in conjunction with tblastn it was used to identify homologs in the prokaryotic genomes.

Molecular Modeling

The serpin from *P. aerophilum* was modeled using the MODELLER program (Sali and Blundell 1993) within the QUANTA package (Accelrys, San Diego, Calif.) using the structure of antithrombin (PDB accession, 1azx) as a template. The alignment between the two proteins was obtained using PSI-BLAST and manually adjusted to take account of elements of secondary structure. Where necessary, manual side-chain refinement was performed using CHARMm. A Ramachandran plot revealed that all residues in the model were in allowed conformations.

Table 1
Serpins in Prokaryotic Genomes

Organism	Accession[a]	Bases	Expect[b]	Id %[c]	Cons.[d]	Source
Bacteria						
Dehalococcoides ethenogenes	(TIGR_61435 Contig 6423)	25100–26161	6×10^{-43}	29	45	TIGR
Anabaena sp. PCC 7120	NC_003272	901827–902951	1×10^{-41}	27	47	Kaneko et al. (2001)
Nostoc punctiforme	(DOE_63737 Contig 423)	2051–3160	5×10^{-39}	27	47	JGI
Desulfitobacterium hafniense	(DOE_49338 Contig 3204)	1879–2967	1×10^{-33}	26	47	JGI
Thermoanaerobacter tengcongensis	NC_003869	1519327–1520598	3×10^{-24}	24	41	Bao et al. (2002)
Thermobifida fusca	(DOE_2021 Contig 64)	125510–124484	4×10^{-23}	25	40	JGI
Ruminococcus albus	(TIGR_1264 Contig 115)	7413–6391	6×10^{-22}	23	38	TIGR
Archaea						
Methanosarcina acetivorans 1	NC_003552	2762806–2764086	3×10^{-39}	28	43	Galagan et al. (2002)
Methanosarcina acetivorans 3	NC_003552	4187242–4188516	4×10^{-39}	29	43	Galagan et al. (2002)
Methanosarcina mazei	AE013514	6329–7609	5×10^{-36}	28	44	Doppenmeier et al. (2002)
Methanosarcina acetivorans 2[e]	NC_003552	3244418–3243145	9×10^{-32}	27	43	Galagan et al. (2002)
Pyrobaculum aerophilum[f]	NC_003364	23237–24355	3×10^{-14}	23	32	Fitz-Gibbon et al. (2002)

[a] Parenthesis indicate internal identifiers to unfinished genomes that do not appear in GenBank.

[b] Expect value from a tblastn search using human α_1-antitrypsin.

[c] Percent identity when aligned with human α_1-antitrypsin.

[d] Number of residues identical to the 51 highly conserved positions in the serpin superfamily (Irving et al. 2000).

[e] This is combination of two predicted open reading frames, divided by a stop codon N-terminal to sheet 5A. If this stop codon is not a sequencing artifact, this protein may either adopt an alternative structure or may be a pseudogene.

[f] The predicted gene presented in the GenBank database encodes a protein that lacks the A-helix. Analyses of ''codons'' upstream of the predicted start site, with reference to conserved residues in the serpin family, strongly suggest that this protein at minimum retains the A-helix, as shown in Figure 1.

Results

A BLAST search of the microbial BLAST database using the protein sequence of α_1-antitrypsin identified seven bacterial serpin-like sequences and five archaeal sequences, with significant expect scores (table 1). These putative proteins were aligned with human serpins α_1-antitrypsin and antithrombin, the tobacco hornworm serpin 1K, and chicken ovalbumin (fig. 1). All 12 contain between 23% and 29% sequence similarity with antitrypsin. Although the lower of these values lie within the ''twilight zone'' of protein sequence comparison, all exhibit the highly conserved motifs characteristic of the serpin superfamily. In particular, the bacterial serpins and most of the archaeal serpins possess 38–47 of the 51 residues strictly conserved in >70% of serpin family members (table 1 and fig. 1). This falls well within the observed range for other serpin family members (on average, 43.5 ± 5.9 residues) and indicates that the sequences encode functional serpins. The serpin from *P. aerophilum* has 32 residues consistent with conserved serpin positions and thus represents the most ''diverged'' prokaryotic sequence. This is likely to reflect the extreme environment that the organism occupies: deep thermal vents with temperatures that can rise above 100°C (Volkl et al. 1993).

We sought to determine whether the presence of serpins in the genomes of these organisms correlates with an endogenous enzyme target. Some inhibition by serpins has been shown for A1 (pepsin), C1 (papain), C14 (caspase), C25 (gingipain), S1 (trypsin), and S8 (subtilisin) proteinase families. A BLAST search revealed that eight of the prokaryotic genomes (with the exception of *D. ethenogenes* and *R. albus*) contained subtilisin-like sequences, and all save *Methanosarcina* spp. contained homologs of the trypsin-like enzyme HtrA. Caspase-like enzymes, a fold that includes the gingipains, have been found in *Anabaena*, *N. punctiforme*, and *D. ethenogenes* species (Aravind and Koonin 2002; Koonin and Aravind 2002). None of the other proteinase families considered were found. It is clear, however, that the presence of these enzymes does not correlate exclusively with presence of serpins. Of the prokaryotic genomes lacking serpins, subtilisin-like enzymes were present in ~20%, trypsin-like enzymes in ~60%, and caspase-like enzymes are also found elsewhere (Koonin and Aravind 2002). A similar search using all proteinase classes in the MEROPS database (as on December 17, 2001) failed to identify other types present exclusively in serpin-bearing genomes.

It has been noted that coexpressed genes are frequently situated as contiguous clusters within an operon. But an analysis of regions adjacent to the serpin sequences did not reveal, from one organism to the next, a consistent pattern of neighboring genes that might indicate the presence of a conserved operon. Five of the sequences were taken from unassembled contig data; thus whether they are present in transposable elements or plasmids is not yet clear. The genes from *Anabaena*

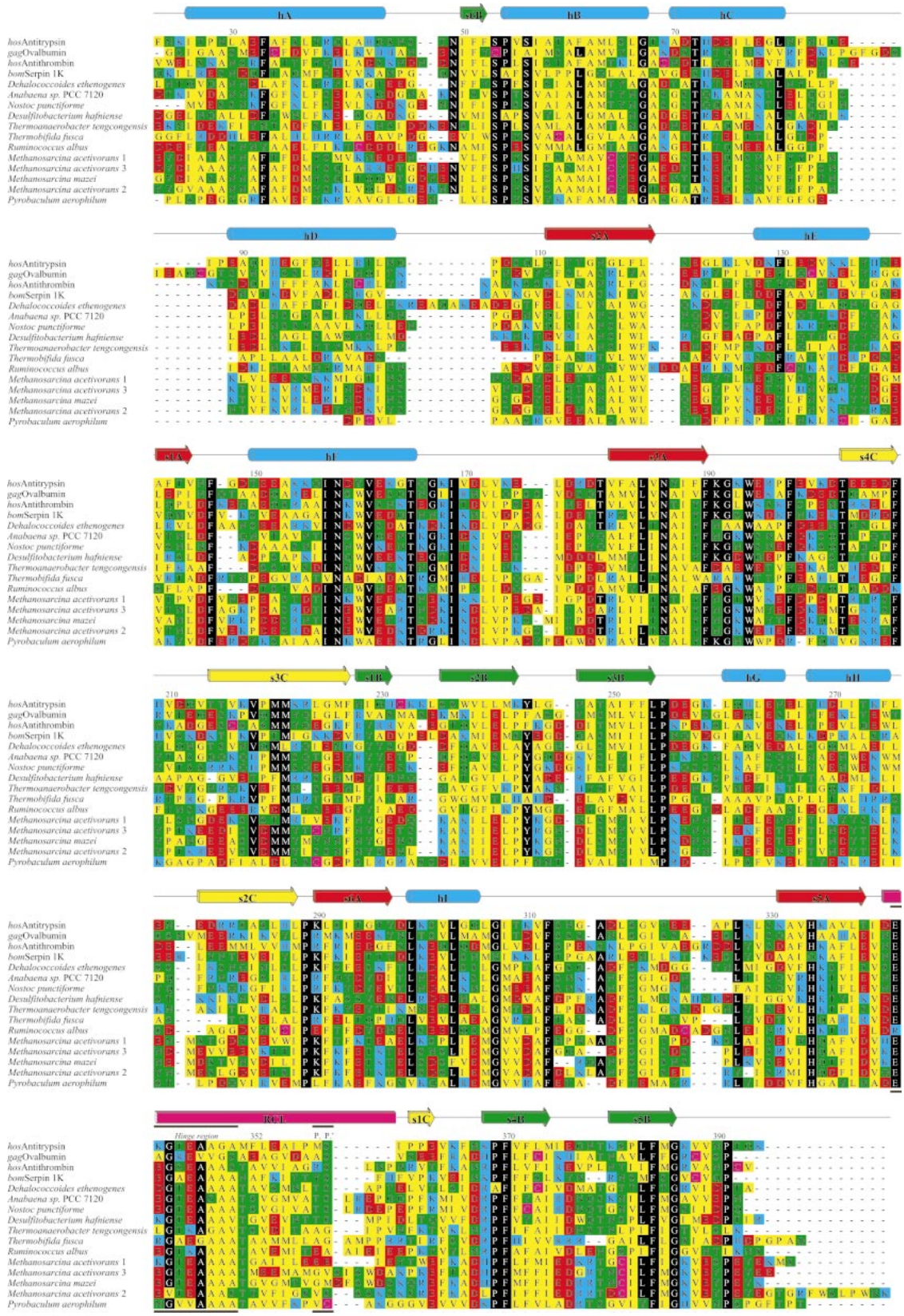

FIG. 1.—Alignment between the prokaryotic serpins and selected members of the serpin superfamily. Each bacterial serpin contains a 3–71 residue N-terminal sequence not shown in the alignment because it is outside the serpin core. These N-terminal extensions share no significant sequence similarity with any known domain. Residues are colored according to type: polar uncharged (green), acidic (red), basic (blue), and nonpolar (yellow). Those residues which are identical to residues conserved in >70% of the superfamily (Irving et al. 2000) appear in white-on-black. Elements of consensus secondary structure are shown at the top. The ''hinge region'' and P_1-P_1' residues are indicated by thick lines. Numbering is according to human α_1-antitrypsin.

sp. PCC 7120, *P. aerophilum*, *T. tengcongensis*, *M. mazei*, and *M. acetivorans* are localized to the single chromosome of each organism.

Recognizable serpin sequences are not ubiquitous or even widely distributed in the prokaryotic kingdom. Serpins appear to be absent from all other sequenced archaeal organisms (13 finished and two unfinished), from proteobacteria (26 finished and 43 unfinished), and from several other branches less well represented in genome sequencing efforts. Furthermore, one finished and four unfinished cyanobacterial genomes (including *N. punctiforme*) and 21 complete firmicute genomes (a family which includes *T. fusca*) also lack serpins.

Discussion

Database searching identified putative serpins in seven species of bacteria and three archaeal species. It is likely that all 12 are functional as proteinase inhibitors because all possess the sequence of small residues within the hinge region, key secondary structural elements, and many of the conserved residues from the family as a whole.

The actinomycete *T. fusca* (formerly *Thermomonospora fusca*) is a thermophilic soil bacterium with an optimal growth temperature of 55°C that plays an important role in the degradation of plant detritus, for example in compost heaps (Crawford 1975). A related organism, the firmicute *T. tengcongensis*, grows at 75°C and is found in hot springs (Xue et al. 2001). The crenarchaeon *P. aerophilum*, a member of which has been obtained from a boiling marine water hole, is a hyperthermophile with an optimal growth temperature of 100°C (Fitz-Gibbon et al. 2002). The existence of serpins in thermophilic organisms is of particular interest because serpins are metastable in their native state and susceptible to heat-induced polymerization (Lomas et al. 1992). Thermophilic organisms' euryarchaeote neighbors, *M. acetivorans* and *M. mazei*, are methane-producing organisms that grow at moderate temperatures (35–40°C) and were first isolated from marine sediments (Sowers, Baron, and Ferry 1984). *Desulfitobacterium hafniense* is a gram-positive, endospore-forming, strictly anaerobic bacterium initially isolated from municipal sludge that is capable of dechlorinating both aromatic and alkyl chlorinated compounds (for a review see El-Fantroussi, Naveau, and Agathos 1998). The filamentous nitrogen-fixing cyanobacterium *N. punctiforme* is able to form symbiotic relationships with a variety of terrestrial plants. Furthermore this organism is able to form a partnership with an obligate symbiotic fungus (Zygomycotina) to form the organism *Geosiphon pyriforme*, the only known example of endocytobiosis (intracellular association of two cells) between a fungus and cyanobacteria (Gehrig, Schussler, and Kluge 1996). *Dehalococcoides ethenogenes* is a eubacterium capable of reducing tetrachloroethene to ethane (Maymo-Gatell et al. 1997). *Dehalococcoides ethenogenes*, along with strain CBDB1 and several uncultivated bacteria, forms part of a clade phylogenetically removed from other bacterial families (Adrian et al. 2000). Finally, *R. albus* is a gram-positive, cellulolytic anaerobe that inhabits the gut of herbivores (Leatherwood 1965). Thus serpins appear to be present in a wide variety of prokaryotes that live in diverse environments.

Bootstrapped neighbor-joining trees were constructed on the basis of the alignment of the prokaryotic serpins with other members of the serpin superfamily (identified in Irving et al. 2000). The majority consensus tree revealed three strictly conserved associations: the two cyanobacteria, *N. punctiforme* and *Anabaena* sp. PCC 7120, clustered together in 100% of the time; the firmicutes *D. hafniense*, *R. albans* and *T. tengcongensis* coincided in 100% of trees; and the euryarchaeote serpins from *Methanosarcina* spp. also formed a clade with 100% support (fig. 2). These associations are not unexpected: they reflect the close evolutionary origin of these prokaryotic species. The comparison consensus method, which identifies underlying relationships that may be obscured by other poorly resolved species, revealed a significant (100%) association between the cyanobacterial and firmicute serpins. It was not possible from these analyses to identify a well-supported relationship between these five bacterial serpins and those from *T. fusca*, *D. ethenogenes*, *P. aerophilum*, the *Methanosarcina* clade or any other subfamily. These five sequences therefore represent a novel bacterial clade in the serpin superfamily, the four *Methanosarcina* proteins form an archaeal clade, and the other three serpins are, at present, ''orphans'' (fig. 2).

Using the tree division method, which quantifies the association between members of the bootstrapped phylogenetic trees and certain prenominated sequences, the following question was asked: do the bacterial sequences associate more closely with (1) vertebrate, (2) arthropod, (3) nematode, or (4) plant serpins? The most highly supported association, between the *T. fusca* serpin and plant sequences, was found in 82.7% of bootstrap trees, suggestive but below the 95% significance threshold of the method. This relationship, although interesting in light of the environmental context of the bacterium, is therefore by no means conclusive. The serpins from cyanobacteria-firmicutes associated more closely with nematode sequences (74.4%) and then with plant sequences (15.4%). Because these associations are fairly strongly biased toward two different lineages, they suggest the possibility of lateral gene transfer.

Of 77 finished genomes, 67 did not contain any identifiable serpin homolog. Three scenarios seem feasible:

1. Serpins have an ancient origin deep in the prokaryotic tree, but most, because of a tendency to rapidly shed unnecessary genes from their genomes (Makarova et al. 1999; Makarova, Ponomarev, and Koonin 2001), have lost serpin homologs. This hypothesis is supported in part by the markedly different environments in which some of these organisms exist, which would make lateral transfer of the serpin genes less likely.
2. Serpins have a relatively recent origin within a subset of the prokaryotic kingdom and have been passed

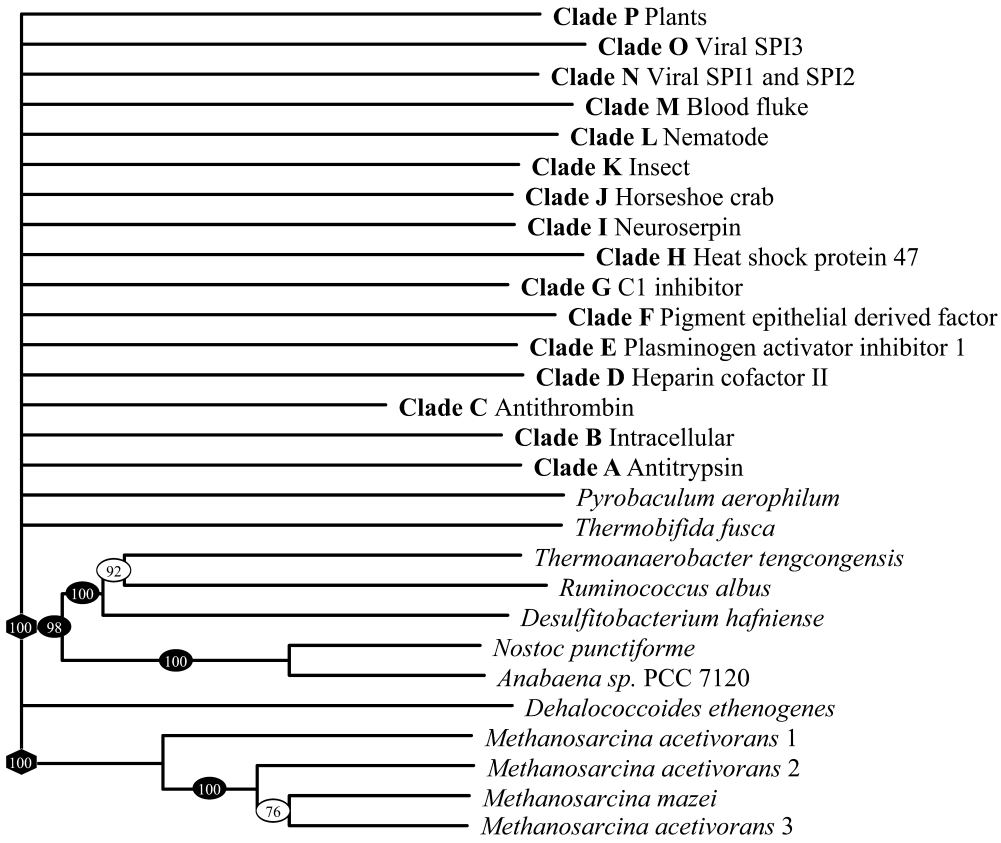

FIG. 2.—Tree illustrating the deduced evolutionary relationships between the bacterial serpins. Conventional bootstrap values are indicated by ovals; the rectangular symbol at the root of the tree indicates the association was determined using the comparison consensus method (Irving et al. 2000); and the hexagon indicates the relationship was strictly conserved in all bootstrap trees. The polytomy at the base of the tree indicates that some of the relationships could not be deduced with confidence. The 219 eukaryotic serpins included in the analysis are summarized by their respective clade designations (Irving et al. 2000). Branch lengths were determined using CLUSTALW (Higgins, Thompson, and Gibson 1996).

through lateral transfer to other prokaryotes and eukaryotes as well.

3. Serpins have an origin within eukaryotes and have been passed by lateral transfer to prokaryotes. The phylogenetic data shows some support for this hypothesis.

The latter two scenarios seem more favorable in light of the sporadic distribution of serpin genes in prokaryotic genomes; however, there is no evidence that strongly favors one over the other. With the elucidation of further sequences of prokaryotic organisms living at the boundaries of life, it should be possible to elucidate the evolutionary origin of serpins with greater certainty.

Sequence alignment of the 12 serpins with serpins of known structure reveals that the bacterial sequences, and all but one of the archaeal sequences, are predicted to contain the full complement of elements of secondary structure (fig. 1). This is in contrast to viral serpins, several of which contain deletion of entire secondary structure elements (Renatus et al. 2000; Simonovic, Gettins, and Volz 2000; Guerin et al. 2001). As with the viral serpins, the protein from the archaeon *P. aerophilum* appears to have lost its D-helix; however, we predict that *P. aerophilum* retains all other structural elements. All 12 prokaryotic serpins contain the pattern of small amino acids in the hinge region of the reactive center loop (RCL) (see fig. 1), characteristic of serpins that function as proteinase inhibitors; hence we suggest that the most likely function of these proteins is the inhibition of proteinases.

One striking feature of four of the bacterial serpins (the two *Nostoc* species, *D. ethenogenes* and *D. hafniense*) is that the P_1 and P_1' residues (refer to abbreviations for P_n notation) are identical (T-S). The RCL has been noted to be variable in duplicated serpin genes undergoing functional diversification (Inglis et al. 1991; Kaiserman et al. 2002), and this is indicative of negative (purifying) selection. Two observations further support this hypothesis. First, these serpins are substantially different from one another overall, with 34%–69% identity. Second, three of the sequences share a strong evolutionary relationship (*Nostoc* spp. and *D. hafniense*; see fig. 2), but *R. albus* and *T. tengcongensis*, which are also predicted to be part of this clade, do not have a T-S at the P_1-P_1'. Furthermore, *D. ethenogenes* (which does not share a significant evolutionary relationship with *Nostoc* spp. or *D. hafniense*) does have a T-S at this site. It therefore appears likely that the four predicted inhibitors from *Nostoc* spp., *D. hafniense*, and *D. ethenogenes* are under selective pressure to target proteinases of similar

or identical specificity. Similarly, *T. fusca* and *T. tengcongensis* both have the residues A-G; *R. albus* and *M. acetivorans* 1 share a P_1E; and *P. aerophilum* and *M. acetivorans* share P_1V.

The sequences A-G and T-S do not fit the substrate profile of any known serine proteinase. But examination of serpins that are known to inhibit papain-like cysteine proteinases reveal that these inhibitors often contain small residues such as Thr at P_1 (Schick et al. 1998; Irving et al. 2002). Furthermore, the P_2 position (usually of greater importance in targeting a serpin to a cysteine proteinase) is usually occupied by a large hydrophobic residue that interacts with the S_2 specificity pocket. The inhibitors from *T. fusca* and *D. ethenogenes* both have a P_2 leucine, and we hypothesize that these serpins may be able to inhibit papain-like cysteine proteinases.

The only serpins known to have an acidic P_1 residue are the granzyme B inhibitor PI-9 (Sun et al. 2001) and the caspase inhibitor crmA (Ray et al. 1992). Nevertheless, it seems unlikely that the *R. albus* serpin (P_1-P_1' sequence E-A) would interact with granzyme B or caspase-1 in a physiological context such as the milieu of the rumen. We note that several bacteria have been shown to express enzymes that can cleave substrates with acidic P_1 residues (Barbosa, Saldanha, and Garratt 1996), including the V8 serine proteinase of the bacterium *Staphylococcus aureus*, which has a homolog in *R. albus*.

Methanosarcina acetivorans is unique among the prokaryotes in that it possesses three serpin genes, each of which has a unique P_1-P_1' sequence, suggesting that each targets a different enzyme. *Methanosarcina mazei* appears to have inherited one of these inhibitors. The close predicted evolutionary relationship between *M. acetivorans* serpin 3 and the *M. mazei* serpin and the similarity in their RCL sequence (P_1-P_1' of G-V and G-M) suggests the two may interact with the same class of proteinase.

We were unable to find a direct correlation between any of the current proteinase family classifications and serpins in the prokaryotic genomes. Therefore, it remains to be determined whether the serpins have evolved to inhibit endogenous proteinases or to target proteinases in the local environment.

The presence of a serpin in *P. aerophilum* is particularly intriguing because its optimum growth temperature (100°C) is well in excess of the temperature that a typical serpin would be expected to remain in the active inhibitory conformation. Numerous studies have demonstrated that serpins are susceptible to heat-induced polymerization, as a consequence of the metastability of the native state (Lomas et al. 1992). This metastability is crucial for serpin inhibitory function: conformational change is required to trap the target proteinase in a distorted, inactive state (Huntington, Read, and Carrell 2000). Certain members of the serpin family do demonstrate enhanced stability—for example, the noninhibitory serpin ovalbumin denatures at 73°C (Dong et al. 2000). Therefore, while it is conceivable that the serpin from *T. fusca* serpin may remain active at 55°C through a dramatic increase in stability of the native form, we suggest that the serpin from *P. aerophilum* must demonstrate some special feature in order for it to use conformational change for inhibition of target proteinases at 100°C. Although it is possible that the *P. aerophilum* protein no longer undergoes conformational change to inhibit target proteinases, the presence of a typical inhibitory ''hinge'' in this serpin suggests that this is not the case. To investigate the *P. aerophilum* serpin further, we built a molecular model using the structure of antithrombin as a template. These data reveal that the predicted *P. aerophilum* serpin is able to adopt the serpin fold and that all essential elements of the serpin ''core'' are present (fig. 3). We predict that the *P. aerophilum* serpin lacks the D-helix, however, this is not unprecedented—the X-ray crystal structure of the viral serpin crmA reveals that the D-helix can be ''lost'' without disrupting the serpin fold (Renatus et al. 2000; Simonovic, Gettins, and Volz 2000). In comparison with typical serpins, the *P. aerophilum* serpin is also predicted to contain a significant insertion at the base of the E-helix. Most strikingly, however, we note that the RCL of the *P. aerophilum* serpin contains a cysteine residue at the P_1' position that we predict would be able to form a disulfide bond with a second cysteine on strand s3C of the C β-sheet. Stabilizing disulfides have been noted in the crystal structure of an intracellular *P. aerophilum* enzyme, adenylosuccinate lyase (Toth et al. 2000). Such an interaction would be predicted to ''tie down'' the RCL and prevent inappropriate conformational change (such as polymerization). In a study based on antitrypsin, in which a disulfide bond was introduced between the RCL and strand s1C of the C β-sheet, the presence of a covalent linkage prevented the serpin from polymerizing (Chang et al. 1997). It was hypothesized

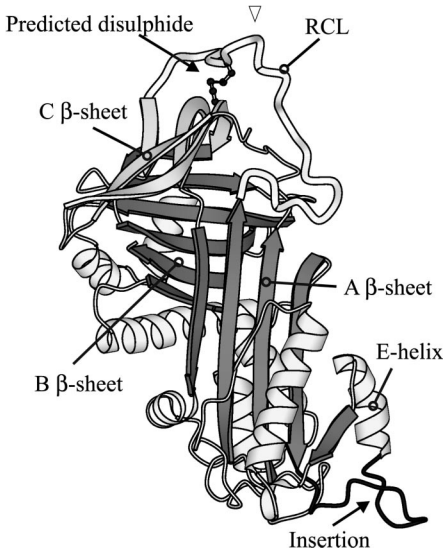

FIG. 3.—Molecular model of the serpin from *P. aerophilum*, indicating the position of the predicted disulfide bond linking the RCL to the C β-sheet. The RCL appears at the top of the molecule, highlighted in black; the cysteine residues are shown in ball-and-stick; and the insertion near the E-helix appears in black near the base of the molecule. Selected elements of secondary structure are labeled. The site in the RCL expected to be cleaved by a target proteinase is indicated by a triangle.

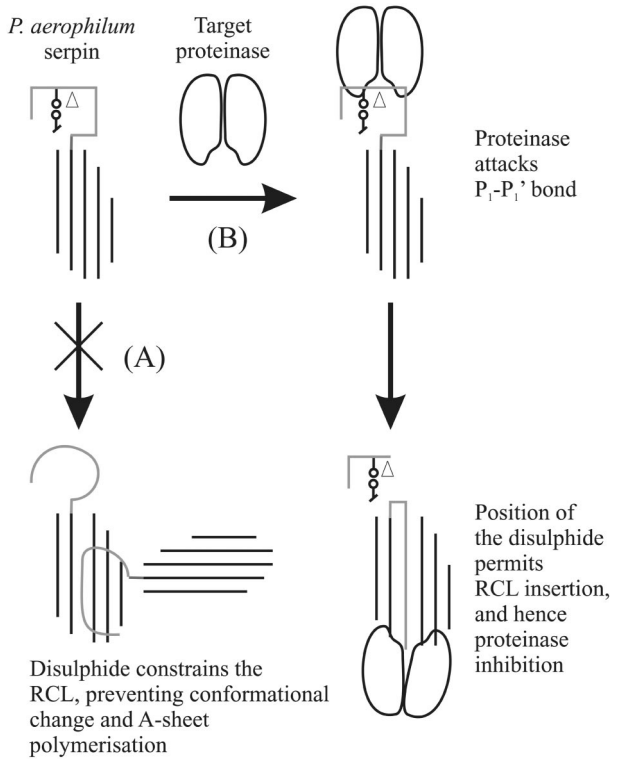

FIG. 4.—A schematic representation of the proposed influence of the RCL–C β-sheet disulfide bond on polymerization, and the serpin inhibitory mechanism. The RCL is shown as a gray line, with the cysteine residues represented by a line-and-circle and the site of cleavage by an empty triangle. The main A β-sheet of the serpin is illustrated by black lines, and the two lobes represent the target proteinase molecule. (*A*) According to the current model of loop-A sheet polymerization (or ''domain-swapping''), the RCL undergoes conformational change and is integrated as a strand into the A β-sheet of another molecule. Consistent with experimental evidence (Chang et al. 1997), the presence of a disulfide would be expected to prevent this interaction by constraining the RCL of *P. aerophilum* serpin. (*B*) The disulfide bond is C-terminal to the scissile bond cleaved by a target proteinase. Consequently, upon cleavage, the RCL is still able to incorporate into the A β-sheet, as seen with the standard serpin mechanism of inhibition (Huntington, Read, and Carrell 2000).

that the introduction of this disulfide bond prevented polymerization by restricting conformational change in the RCL and the first strand of the C β-sheet. We predict that a similar situation exists in the *P. aerophilum* serpin (fig. 4*A*). A related study showed that the presence of the introduced disulfide bond in antitrypsin did not affect inhibitory activity (Hopkins et al. 1997). Similarly, the predicted disulfide bond in *P. aerophilum* serpin would not be expected to affect the inhibitory mechanism because the RCL would still be able to rapidly insert into the A β-sheet after cleavage at the P_1-P_1' (fig. 4*B*). Thus we predict that nature may have used disulfide bonds as a method of stabilizing serpins in the most primitive and extreme environments.

The discovery of serpins in both prokaryotic kingdoms has important implications for the evolution of the serpin superfamily, which was previously believed to be restricted to higher eukaryotic organisms and their viruses. It is not yet clear whether the presence in prokaryotes is the product of gene inheritance or lateral transfer. In time, as sequencing efforts progress, this question may be answered. Nevertheless, these proteins should present useful information on adaptation of the serpin scaffold to the extreme environments in which many prokaryotic organisms live.

Acknowledgments

We thank the Australian Research Council and the National Health and Medical Research Council of Australia for support. J.C.W. is a NHMRC Senior Research Fellow and Monash University Logan Fellow. A.M.L. is supported by the Wellcome Trust. Preliminary data were obtained from the DOE Joint Genome Institute (JGI) at http://www.jgi.doe.gov and The Institute for Genomic Research (TIGR) at http://www.tigr.org, from projects funded by the DOE (with the exception of *R. albus*, funded by the United States Department of Agriculture). The sequence data from the JGI have been provided freely by the US DOE Joint Genome Institute for use in this publication only.

LITERATURE CITED

ADACHI, J., and M. HASEGAWA. 1996. MOLPHY: programs for molecular phylogenetics. Version 2.3. Institute of Statistical Mathematics, Tokyo.

ADRIAN, L., U. SZEWZYK, J. WECKE, and H. GORISCH. 2000. Bacterial dehalorespiration with chlorinated benzenes. Nature **408**:580–583.

ARAVIND, L., and E. V. KOONIN. 2002. Classification of the caspase-hemoglobinase fold: detection of new families and implications for the origin of the eukaryotic separins. Proteins **46**:355–367.

BAO, Q., Y. TIAN, W. LI et al. (21 co-authors). 2002. A complete sequence of the T. tengcongensis genome. Genome Res. **12**:689–700.

BARBOSA, J. A., J. W. SALDANHA, and R. C. GARRATT. 1996. Novel features of serine protease active sites and specificity pockets: sequence analysis and modelling studies of glutamate-specific endopeptidases and epidermolytic toxins. Protein Eng. **9**:591–601.

CHANG, W. S., J. C. WHISSTOCK, P. C. HOPKINS, A. M. LESK, R. W. CARRELL, and M. R. WARDELL. 1997. Importance of the release of strand 1C to the polymerization mechanism of inhibitory serpins. Protein Sci. **6**:89–98.

CRAWFORD, D. L. 1975. Cultural, morphological, and physiological characteristics of Thermomonospora fusca (strain 190Th). Can. J. Microbiol. **21**:1842–1848.

DAHLEN, J. R., D. C. FOSTER, and W. KISIEL. 1997. Human proteinase inhibitor 9 (PI9) is a potent inhibitor of subtilisin A. Biochem. Biophys. Res. Commun. **238**:329–333.

DEPPENMEIER, U., A. JOHANN, T. HARTSCH et al. (22 co-authors). 2002. The genome of Methanosarcina mazei: evidence for lateral gene transfer between bacteria and archaea. J. Mol. Microbiol. Biotechnol. **4**:453–461.

DONG, A., J. D. MEYER, J. L. BROWN, M. C. MANNING, and J. F. CARPENTER. 2000. Comparative fourier transform infrared and circular dichroism spectroscopic analysis of alpha1-proteinase inhibitor and ovalbumin in aqueous solution. Arch. Biochem. Biophys. **383**:148–155.

EL-FANTROUSSI, S., H. NAVEAU, and S. N. AGATHOS. 1998. Anaerobic dechlorinating bacteria. Biotechnol. Prog. **14**: 167–188.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783–791.

FITZ-GIBBON, S. T., H. LADNER, U. J. KIM, K. O. STETTER, M. I. SIMON, and J. H. MILLER. 2002. Genome sequence of the hyperthermophilic crenarchaeon Pyrobaculum aerophilum. Proc. Natl. Acad. Sci. USA **99**:984–989.

GALAGAN, J. E., C. NUSBAUM, A. ROY et al. (55 co-authors). 2002. The genome of M. acetivorans reveals extensive metabolic and physiological diversity. Genome Res. **12**:532–542.

GEHRIG, H., A. SCHUSSLER, and M. KLUGE. 1996. Geosiphon pyriforme, a fungus forming endocytobiosis with Nostoc (cyanobacteria), is an ancestral member of the Glomales: evidence by SSU rRNA analysis. J. Mol. Evol. **43**:71–81.

GUERIN, J. L., J. GELFI, C. CAMUS, M. DELVERDIER, J. C. WHISSTOCK, M. F. AMARDEIHL, R. PY, S. BERTAGNOLI, and F. MESSUD-PETIT. 2001. Characterization and functional analysis of Serp3: a novel myxoma virus-encoded serpin involved in virulence. J. Gen. Virol. **82**:1407–1417.

HIGGINS, D. G., J. D. THOMPSON, and T. J. GIBSON. 1996. Using CLUSTAL for multiple sequence alignments. Methods Enzymol. **266**:383–402.

HOPKINS, P. C., R. W. CARRELL, and S. R. STONE. 1993. Effects of mutations in the hinge region of serpins. Biochemistry **32**:7650–7657.

HOPKINS, P. C., W. S. CHANG, M. R. WARDELL, and S. R. STONE. 1997. Inhibitory mechanism of serpins. Mobility of the C-terminal region of the reactive-site loop. J. Biol. Chem. **272**:3905–3909.

HUNTINGTON, J. A., R. J. READ, and R. W. CARRELL. 2000. Structure of a serpin-protease complex shows inhibition by deformation. Nature **407**:923–926.

INGLIS, J. D., M. LEE, D. R. DAVIDSON, and R. E. HILL. 1991. Isolation of two cDNAs encoding novel alpha 1-antichymotrypsin-like proteins in a murine chondrocytic cell line. Gene **106**:213–220.

IRVING, J. A., R. N. PIKE, A. M. LESK, and J. C. WHISSTOCK. 2000. Phylogeny of the serpin superfamily implications of patterns of amino acid conservation for structure and function. Genome Res. **10**:1845–1864.

IRVING, J. A., S. S. SHUSHANOV, R. N. PIKE, E. Y. POPOVA, D. BROMME, T. H. COETZER, S. P. BOTTOMLEY, I. A. BOULYNKO, S. A. GRIGORYEV, and J. C. WHISSTOCK. 2002. Inhibitory activity of a heterochromatin-associated serpin (MENT) against papain-like cysteine proteinases affects chromatin structure and blocks cell proliferation. J. Biol. Chem. **277**:13192–13201.

KAISERMAN, D., S. KNAGGS, K. L. SCARFF et al. (13 co-authors). 2002. Comparison of human chromosome 6p25 with mouse chromosome 13 reveals a greatly expanded ov-serpin gene repertoire in the mouse. Genomics **79**:349–362.

KANEKO, T. Y. N., C. P. WOLK et al. (22 co-authors). 2001. Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. DNA Res. **8**:203–215.

KOMIYAMA, T., H. GRON, P. A. PEMBERTON, and G. S. SALVESEN. 1996. Interaction of subtilisins with serpins. Protein Sci. **5**:874–882.

KOONIN, E. V., and L. ARAVIND. 2002. Origin and evolution of eukaryotic apoptosis: the bacterial connection. Cell Death Differ. **9**:394–404.

LEATHERWOOD, J. M. 1965. Cellulase from Ruminococcus albus and mixed rumen microorganisms. Appl. Microbiol. **13**:771–775.

LIPINSKA, B., M. ZYLICZ, and C. GEORGOPOULOS. 1990. The HtrA (DegP) protein, essential for Escherichia coli survival at high temperatures, is an endopeptidase. J. Bacteriol. **172**:1791–1797.

LOMAS, D. A., D. L. EVANS, J. T. FINCH, and R. W. CARRELL. 1992. The mechanism of Z alpha 1-antitrypsin accumulation in the liver. Nature **357**:605–607.

MAKAROVA, K. S., L. ARAVIND, M. Y. GALPERIN, N. V. GRISHIN, R. L. TATUSOV, Y. I. WOLF, and E. V. KOONIN. 1999. Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. Genome Res. **9**:608–628.

MAKAROVA, K. S., V. A. PONOMAREV, and E. V. KOONIN. 2001. Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. Genome Biol. **2**:research0033.

MAYMO-GATELL, X., Y. CHIEN, J. M. GOSSETT, and S. H. ZINDER. 1997. Isolation of a bacterium that reductively dechlorinates tetrachloroethene to ethene. Science **276**:1568–1571.

NAKAI, A., M. SATOH, K. HIRAYOSHI, and K. NAGATA. 1992. Involvement of the stress protein HSP47 in procollagen processing in the endoplasmic reticulum. J. Cell Biol. **117**:903–914.

RAWLINGS, N. D., E. O'BRIEN, and A. J. BARRETT. 2002. MEROPS: the protease database. Nucleic Acids Res. **30**:343–346.

RAY, C. A., R. A. BLACK, S. R. KRONHEIM, T. A. GREENSTREET, P. R. SLEATH, G. S. SALVESEN, and D. J. PICKUP. 1992. Viral inhibition of inflammation: cowpox virus encodes an inhibitor of the interleukin-1 beta converting enzyme. Cell **69**:597–604.

RENATUS, M., Q. ZHOU, H. R. STENNICKE, S. J. SNIPAS, D. TURK, L. A. BANKSTON, R. C. LIDDINGTON, and G. S. SALVESEN. 2000. Crystal structure of the apoptotic suppressor CrmA in its cleaved form. Structure **8**:789–797.

SALI, A., and T. L. BLUNDELL. 1993. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. **234**:779–815.

SCHECTER, I., and A. BERGER. 1967. On the size of the active site in proteases. I. Papain. Biochem. Biophys. Res. Commun. **27**:157–162.

SCHICK, C., P. A. PEMBERTON, G. P. SHI, Y. KAMACHI, S. CATALTEPE, A. J. BARTUSKI, E. R. GORNSTEIN, D. BROMME, H. A. CHAPMAN, and G. A. SILVERMAN. 1998. Cross-class inhibition of the cysteine proteases cathepsins K, L, and S by the serpin squamous cell carcinoma antigen 1: a kinetic analysis. Biochemistry **37**:5258–5266.

SILVERMAN, G. A., P. I. BIRD, R. W. CARRELL et al. (15 co-authors). 2001. The serpins are an expanding superfamily of structurally similar but functionally diverse proteins. Evolution, mechanism of inhibition, novel functions, and a revised nomenclature. J. Biol. Chem. **276**:33293–33296.

SIMONOVIC, M., P. G. W. GETTINS, and K. VOLZ. 2000. Crystal structure of viral serpin crmA provides insights into its mechanism of cysteine proteinase inhibition. Protein Sci. **9**:1423–1427.

SNIPAS, S. J., H. R. STENNICKE, S. RIEDL, J. POTEMPA, J. TRAVIS, A. J. BARRETT, and G. S. SALVESEN. 2001. Inhibition of distant caspase homologues by natural caspase inhibitors. Biochem. J. **357**:575–580.

SOWERS, K. R., S. F. BARON, and J. G. FERRY. 1984. Methanosarcina acetivorans sp. nov., an acetotrophic methane-producing bacterium isolated from marine sediments. Appl. Environ. Microbiol. **47**:971–978.

STEIN, P. E., and R. W. CARRELL. 1995. What do dysfunctional serpins tell us about molecular mobility and disease? Nat. Struct. Biol. **2**:96–113.

SUN, J., J. C. WHISSTOCK, P. HARRIOTT, B. WALKER, A. NOVAK, P. E. THOMPSON, A. I. SMITH, and P. I. BIRD. 2001. Importance of the P4′ residue in human granzyme B inhibitors and substrates revealed by scanning mutagenesis of the proteinase inhibitor 9 reactive center loop. J. Biol. Chem. **276**:15177–15184.

TOTH, E. A., C. WORBY, J. E. DIXON, E. R. GOEDKEN, S. MARQUSEE, and T. O. YEATES. 2000. The crystal structure of adenylosuccinate lyase from Pyrobaculum aerophilum reveals an intracellular protein with three disulfide bonds. J. Mol. Biol. **301**:433–450.

VOLKL, P., R. HUBER, E. DROBNER, R. RACHEL, S. BURGGRAF, A. TRINCONE, and K. O. STETTER. 1993. Pyrobaculum aerophilum sp. nov., a novel nitrate-reducing hyperthermophilic archaeum. Appl. Environ. Microbiol. **59**:2918–2926.

XUE, Y., Y. XU, Y. LIU, Y. MA, and P. ZHOU. 2001. Thermoanaerobacter tengcongensis sp. nov., a novel anaerobic, saccharolytic, thermophilic bacterium isolated from a hot spring in Tengcong, China. Int. J. Syst. Evol. Microbiol. **51**: 1335–1341.

PEER BORK, reviewing editor

Accepted June 10, 2002