PROTEINS: Structure, Function, and Bioinformatics 67:681–694 (2007)

Evolution of Prokaryotic Subtilases: Genome-Wide Analysis Reveals Novel Subfamilies With Different Catalytic Residues

Roland J. Siezen,[1,2*] **Bernadet Renckens,**[1,2] **and Jos Boekhorst**[1]
[1]*Center for Molecular and Biomolecular Informatics, Radboud University, Nijmegen, the Netherlands*
[2]*NIZO food research, Ede, the Netherlands*

ABSTRACT **Subtilisin-like serine proteases (subtilases) are a very diverse family of serine proteases with low sequence homology, often limited to regions surrounding the three catalytic residues. Starting with different Hidden Markov Models (HMM), based on sequence alignments around the catalytic residues of the S8 family (subtilisins) and S53 family (sedolisins), we iteratively searched all ORFs in the complete genomes of 313 eubacteria and archaea. In 164 genomes we identified a total of 567 ORFs with one or more of the conserved regions with a catalytic residue. The large majority of these contained all three regions around the "classical" catalytic residues of the S8 family (Asp-His-Ser), while 63 proteins were identified as S53 (sedolisin) family members (Glu-Asp-Ser). More than 30 proteins were found to belong to two novel subsets with other evolutionary variations in catalytic residues, and new HMMs were generated to search for them. In one subset the catalytic Asp is replaced by an equivalent Glu (i.e. Glu-His-Ser family). The other subset resembles sedolisins, but the conserved catalytic Asp is not located on the same helix as the nucleophile Glu, but rather on a β-sheet strand in a topologically similar position, as suggested by homology modeling. The Prokaryotic Subtilase Database (www.cmbi.ru.nl/subtilases) provides access to all information on the identified subtilases, the conserved sequence regions, the proposed family subdivision, and the appropriate HMMs to search for them. Over 100 proteins were predicted to be subtilases for the first time by our improved searching methods, thereby improving genome annotation. Proteins 2007;67:681–694.** © 2007 Wiley-Liss, Inc.

Key words: subtilisin; sedolisin; serine protease; genome; archaea; gram-positive bacteria; gram-negative bacteria

INTRODUCTION

Serine peptidases of the SB clan,[1] also known as the subtilase superfamily, are a very diverse family of subtilisin-like serine proteases found in archaea, eubacteria, fungi, yeasts, and higher eukaryotes.[2–5] Prokaryotic subtilases are generally secreted outside the cell, and are mainly known to play a role in either nutrition (providing peptides and amino acids for cell growth) or host invasion (e.g., degradation of host cell–surface receptors or host enzyme inhibitors), such as the C5a peptidase of *Streptococcus pyogenes*.[6] In recent years it has been shown that subtilases are also involved in various precursor processing and maturation reactions, both intracellularly and extracellularly. In prokaryotes, subtilases are known to be maturation proteases for (i) bacteriocins, such as the lantibiotics,[7] (ii) extracellular adhesins, such as filamentous haemagglutinin,[8] and (iii) spore-germination enzymes, such as spore-cortex lytic enzyme of *Clostridium*.[9] Subtilases encoded in conserved ESAT-6 gene clusters in mycobacteria, *Corynebacterium diphtheriae*, and *Strepomyces coelicolor* are postulated to be involved in maturation of secreted T-cell antigens.[10]

Most subtilases have a multi-domain structure consisting of a signal peptide (for translocation), a pro-peptide (for maturation by autoproteolytic cleavage), a protease domain, and frequently one or more additional domains.[2,11,12] Subtilases lacking a signal peptide should remain inside the cell, and most likely play a role in intracellular maturation of other proteins and peptides. Extracellular subtilases can remain attached to the cell wall if they have additional anchoring domains, such as an LPxTG motif for binding to peptidoglycan.[13–15]

The overall sequence identity of the protease domain was known to be low, and until a few years ago it was thought that only the three catalytic residues Asp, His, and Ser were totally conserved while only short segments surrounding these residues showed low conservation throughout the entire family.[2,11] Recently, crystal structure determination of three bacterial sedolisins (or carboxyl serine peptidase) demonstrated that they constitute a novel family S53 of clan SB, with folding very similar to that of subtilisins, in which the catalytic triad has been

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsor: Bio Range program of the Netherlands Bio Informatics Centre (NBIC), funded by BSIK through Netherlands Genomics Initiative (NGI).

*Correspondence to: Roland J. Siezen, NIZO food research, PO Box 20, 6710BA Ede, the Netherlands. E-mail: siezen@cmbi.ru.nl

Received 15 May 2006; Revised 21 August 2006; Accepted 6 October 2006

Published online 8 March 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21290

© 2007 WILEY-LISS, INC.

altered to Glu, Asp, and Ser, and the oxyanion hole Asp replaces Asn, leading to peptidases active at acidic pH, unlike the homologous subtilisins.[16,17] Sedolisins are also widespread in fungi and other eukaryotes[18,19]

In the past few years, complete genome sequences for hundreds of microbial genomes have become available; see for instance the Comprehensive Microbial Resource[20] (http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi). Because of the large sequence diversity among subtilases, including the variation in catalytic residues, identification of new family members is not always straightforward. In fact, only the MEROPS and SCOP databases distinguish between the S8 (subtilisins) and S53 (sedolisins) families, whereas others such as TIGRFAMs, Pfam, Interpro, UniProt, PRINTS, BLOCKs, and PROSITE do not, leading to numerous unidentified or overpredicted subtilases in these databases. To provide better search algorithms to identify subtilases and distinguish between the families, we have now developed and used different Hidden Markov Models (HMMs), based on conserved sequences surrounding the different catalytic residues, to identify all subtilases encoded in prokaryote genomes. Using multiple sequence alignments and homology modeling, we also identified a third subfamily resembling sedolisins with yet another Glu-Asp-Ser catalytic triad, and some evolutionary variants with Glu-His-Ser triads.

METHODS

HMM Searching and Sequence Analysis

The initial set for our search methods consisted of all 45 sequences in the Pfam database[21] alignment of subtilases (PF00082, seed set only). We selected the most conserved regions around the three active site residues Asp (D), His (H), and Ser (H) from the Pfam alignment. The conserved region boundaries are based on the sequence alignment of nearly 200 subtilases.[2] HMMs were built from these three smaller alignments, called (D-H-S)/D-HMM, (D-H-S)/H-HMM, and (D-H-S)/S1-HMM. We used the HMMer package with default settings[22] to build these HMMs, and then searched iteratively against all completed bacterial and archaeal genomes from the NCBI (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html) as of February 2nd, 2006. The 313 genomes searched are listed in Supplementary Table S1. After every search (iteration) the hits with $E < 10^{-03}$ were added to the alignments and new HMMs were made, until no new hits were found below this threshold. Translated open-reading frames (ORFs) with good hits to only a subset of the HMMs (e.g. good hits with (D-H-S)/H-HMM and (D-H-S)/S1-HMM, but no hit with (D-H-S)/D-HMM) were searched for alternative conserved regions using multiple sequence alignments, leading to the identification of additional subtilase families. For instance, members of the sedolisin family (ED-S family) were searched in genomes with a HMM based on a conserved region of 17 residues around the catalytic Glu-x-x-x-Asp (ED) region, called (ED-S)/ED-HMM.

Multiple sequence alignments were created with Clustal W[23,24] or MUSCLE.[25] Phylogenetic trees were constructed using PHMYL.[26]

Prediction of Signal Peptides and Anchors

Prediction of intracellular or extracellular location of a subtilase was based on the (predicted) absence or presence of a signal peptide for sec-dependent translocation,[27] using SignalP 3.0.[28] Carboxy-terminal LPxTG-type anchors were searched with a specific HMM for this motif.[13] Sequences with this motif are cleaved by dedicated sortases resulting in covalent linking of the protein to the bacterial peptidoglycan layer.[14]

Homology Modeling

The three-dimensional structures of subtilisin (PDB code 2SNI) and kumamolysin or KSCP (PDB code 1GTJ) were used as templates of the S8 and S53 families, respectively. Homology modeling of the catalytic domain of selected subtilase variants was performed using 1GTJ as template with "The Whatif/Yasara Twinset" software (www.yasara.com). Models of the E-D-S family include substitutions of catalytic residues Glu32 to Ser, of Ser128 to Asp, and of Asp164 to Asn. Models of the E-H-S family include substitutions of Glu78 to His and of Asp164 to Asn. Optimal rotamer positions for putative catalytic residues were selected.

RESULTS

Genome Searches for Prokaryote Subtilases

Starting with different HMMs, based on sequence alignments around the catalytic residues of the S8 family (subtilisins) and S53 family (sedolisins), we iteratively searched all ORFs in the genomes of over 300 bacteria and archaea. In 164 genomes we identified a total of 567 ORFs with one or more of the conserved regions with a catalytic residue (Table I). The large majority (472) of these identified subtilases contained all three regions around the "classical" catalytic residues Asp, His, and Ser of the S8 family. We will refer to these as the D-H-S family, described in more detail later.

A total of 63 proteins were identified as S53 (sedolisin) family members, based on the combined presence of the two characteristic regions around the Glu-x-x-x-Asp (separated by one helix turn) and Ser catalytic residues. This S53 family, referred to as the ED-S family, is also described in more detail later.

In 32 subtilase hits the catalytic Ser region was identified with the S1-HMM, but other regions around catalytic residues were not identified or scored poorly with the initial HMMs from Pfam. Multiple sequence alignments of these remaining subtilases revealed one very clear subset resembling the S53 family, but with a different conserved Asp residue, here referred to as the E-D-S family. In addition, another subset related to the S8 family was found in which the original Asp is replaced by a Glu catalytic residue (referred to as the E-H-S family). Both new subsets are described later in more detail.

TABLE I. Summary of Subtilases Found with Different HMM Models

Familiy[a]	HMM[b]			Subtilases
D-H-S	D_H_S∼D	D_H_S∼H	D_H_S∼S1	
	1	1	1	438
	0	1	1	4
	1	0	1	9
	1	1	0	5
	0	0	1	11
	1	0	0	5
E-H-S	E_H_S∼E	D_H_S∼H	D_H_S∼S1	
	1	1	1	9
	0	1	1	6
	1	0	1	1
	1	0	0	2
ED-S		ED_S∼ED	D_H_S∼S2	
		1	1	59
		1	0	3
		0	1	1
E-D-S	E_D_S∼E	E_D_S∼D	D_H_S∼S1	
	1	1	1	14
Total				567

[a]The four different familes of subtilases. D-H-S, classical subtilisin family with a catalytic triad consisting of Asp-His-Ser; E-H-S, newly identified family with catalytic residues Glu-His-Ser, whereby the Glu is equivalent to the Asp of the D-H-S family; ED-S, sedolisin family S53 (or serine carboxyl proteinases) with the catalytic residues Glu-Asp-Ser, whereby the Glu and Asp are in the same sequence region; E-D-S, newly identified family with the catalytic residues Glu-Asp-Ser, whereby the Glu and Asp are in different sequence regions. See text for more details.

[b]Presence (1) or absence (0) of identified regions surrounding catalytic residues using different HMMs. For example, D_H_S∼H represents the HMM for the sequence region surrounding the catalytic His in the D-H-S family of subtilases. The large majority of absent motifs is the result of split genes (e.g. leading to two consecutive genes with scores 1-0-0 and 0-1-1) and gene truncations.

Some of the identified subtilase genes were found to contain frame shifts or truncations and hence cannot encode a functional subtilase, although in a few cases this may be the result of incorrect identification of the start codon. The HMMs and all identified subtilases and their predicted properties are listed in the Prokaryote Subtilase Database (http://www.cmbi.ru.nl/subtilases). A list of the number of identified subtilases in all organisms is given in Supplementary Table S2.

D-H-S Family S8 (Subtilisins)

Members of the classical family S8 subtilases (or subtilisin-like serine proteases) have a catalytic triad consisting of Asp32, His64, and Ser221 (numbering of subtilisin) [Fig 1(a)]. In catalysis, Ser221 is the nucleophile and His64 is the general base that accepts the proton from the nucleophilic OH group, while Asp32 stabilizes and orients the general base in the correct position. The sidechain amide of the Asn155 residue contributes to the oxyanion binding site in stabilization of the tetrahedral intermediate. Nearly twenty crystal structures of this SCOP family 52744 are available (http://scop.mrc-lmb.cam.ac.uk/scop/). Two subfamilies are distinguished: the subtilisin S8A subfamily and the kexin S8B subfamily. The latter subfamily is found mostly in eukaryotes. Most members are active at neutral to mildly alkaline pH.

Table I shows that the large majority of prokaryotic subtilases were found to belong to the classical D-H-S family. Most members of this family will be identified using the current subtilase HMMs and motifs in databases such as Pfam (PF00082), Interpro (IPR00209), Prosite (PDOC00125), or PRINTS/BLOCKS (PR00723), but several will still be missed. The new HMMs we have developed iteratively here to find D-H-S family members perform considerably better in this respect, since they are based on a much larger set of sequences, while members of other (sub)families, described later, have been excluded.

ED-S Family S53 (Sedolisins)

The newly identified sedolisin family S53 (or serine carboxyl proteinases) with the subtilase fold has catalytic residues Glu78, Asp82, and Ser278 (numbering of kumamolysin) [Fig. 1(c)]. While the Ser residue remains the nucleophile in sedolisins, the Glu78 residue is in a stereochemically equivalent position to His64 of subtilisin and plays the same role of general base.[29] The Asp residue that orients the general base side chain is in a quite different position, being Asp82 in family S53 (closely following Glu78 in the sequence), in contrast to Asp32 preceding His64 in subtilisin.

Asp164 of the oxyanion binding site, the equivalent of Asn155 in subtilisin, needs to be protonated to function properly, and therefore sedolisins are optimally active at acidic pH.[30,31] Members of this family have been shown to be acid-acting endopeptidases or tripeptidyl peptidases.[18,30,31] Several crystal structures of this SCOP family 52764 are now available (http://scop.mrc-lmb.cam.ac.uk/scop/), e.g. sedolisin from *Pseudomonas* sp. 101,[17] kumamolisin from *Bacillus* novo sp. MN-32,[32,33] and kumamolisin-As from *Alicylobacillus senaiensis* NTAP-1.[16]

Using our new ED-HMM for the Glu-Asp region and an improved HMM for the Ser region of this family (S2-HMM), we have now iteratively identified 63 ED-S family members in prokaryote genomes (Table I), and several others in the NCBI database (Table II). These S53 family proteins are more commonly found in archaea and gram-negative bacteria, with only a few occurrences discovered as yet in gram-positive bacteria. Some organisms appear to have only (or preferably) subtilases of this subfamily, i.e. *Thermoplasma* (*acidophilum*/*volcanium*), *Picrophilus (torridus*), and *Sulfolobu*s (*acidocaldarius*/*solfataricus*/*tokodaii*), which may relate to the very acidic and high temperature environment in which they occur.

The Glu78, Asp82, and Ser278 catalytic residues are found to be invariable in all sequences of the ED-S family. In many cases the original Asp32 is also retained, or sometimes replaced by Glu32 or Thr32 (Table II). Studies

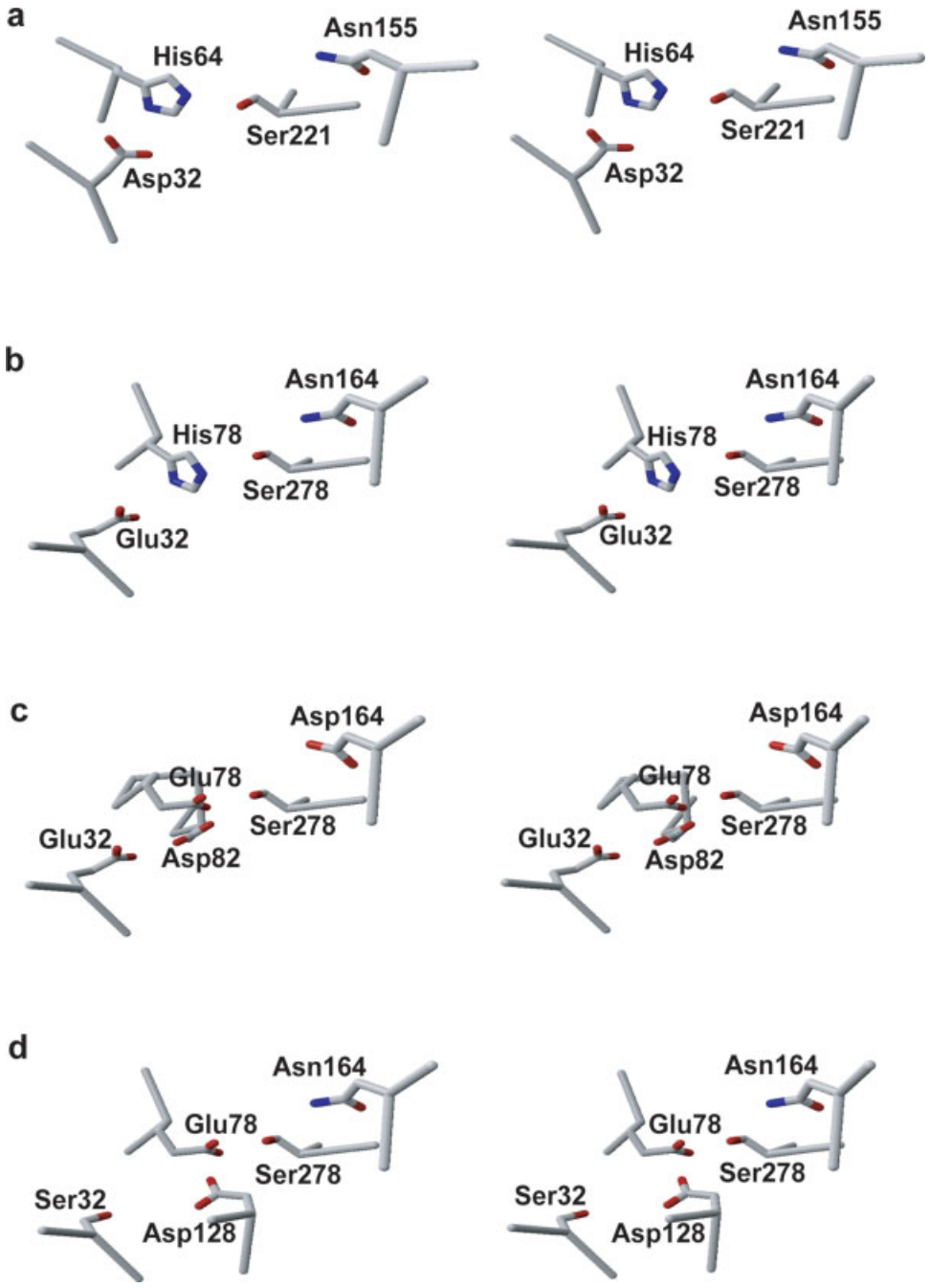

Fig. 1. Stereo views of the catalytic site residues. (**a**) D-H-S family, 3D structure of subtilisin (PDB code 2SNI), (**b**) E-H-S family, homology model derived from kumamolisin (PDB code 1GTJ) by substituting E78 to H78, (**c**) ED-S family, 3D structure of kumamolisin, (**d**) E-D-S family, homology model derived from kumamolisin by substituting E32 to S32, S128 to D128, and D82 to M82 (not shown).

of kumamolisin have shown that additional stabilization of the catalytic residues is created through an extended network of charges and hydrogen bonds via Glu78 and Asp82, including the Glu32-Trp129 pair and several water molecules.[32,33] Therefore, we propose that more variations can occur in the stabilizing hydrogen-bonded network, involving variations in residue 32.

E-D-S Family

A subset of 14 subtilase sequences was found that scored well with the S1-HMM for the region surrounding the catalytic Ser, but did not score well with HMMs for regions surrounding the other catalytic residues in the D-H-S or the ED-S families (Table III). A multiple sequence alignment (Supplementary material Figure S1) shows

TABLE II. The ED.S (sedolisin) Subgroup

Organism	Accession (GI code)	"Normal Asp-region"	Glu-Asp-region	Ser-region
Consensus normal D-H-S subtilases		GKGvtVAViDtGvd-YnHpdL	xxHGthvagiig	sGTSmAaPhvaGvaA
3D structures				
Bacillus novo sp. MN32 (KSCP)	21730221	GQGQCIAIIELGGGYDETSLA	DGEVELDIEVAGALAPG	GGTSAVAPLFAALVA
Pseudomonas sp. (PSCP)	12084517	AANTTVGIITIGGVSQTLQDL	QGEWDLDSQSIVGSAGG	GGTSLASPIFVGLWA
Xanthomonas sp. (XSCP)	1217603	ATNTAVGIITWGSITQTVTDL	NGEWSLDSQDIVGIAGG	GGTSLASPLFVGAFA
Genome hits*				
Bradyrhizobium japonicum usda 110	27375805	GAGQCIAIIELNDIDQKGHPT	DGEVVLDIEVAGAIAPG	GGTSAVAPLMAGLIA
Burkholderia pseudomallei K96243	53719751	ASQTTVGVIMAGDAAPVLRDL	LSEWDMDSQAIVGAAGG	GGTSLAAPIFTGIWA
Burkholderia pseudomallei K96243	53720249	GDGMVVAIVDAYDDPKIESDL	ALEMSLDVEWVHAIAPK	GGTSAGAPQWAALFA
Burkholderia pseudomallei K96243	53722583	GAGQCIAIVELGGGYRPAEIQ	DGEVALDIEIAGAIAPG	GGTSAVAPLWAALVA
Burkholderia pseudomallei K96243	53722755	AANATVGIITIGGVSQALSDL	QGEWDLDSQSIVGAAGG	GGTSLSAPIFTGFWA
Burkholderia pseudomallei K96243	53722994	ATNTTVGIITWGDMTQTIADL	PGEWDLDSQTIIGTSGG	GGTSLASPIFVGGWA
Chromobacterium violaceum ATTC 12472	34497420	AKNGVAGIIAEGNLSQTVADL	IMEWNLDSQTMLAASGG	GGTSLAAPLFSGFWI
Chromobacterium violaceum ATTC 12472	34497423	ASNTTVGIIAEGDLTQTLQDL	VGEWNLDSQDILAAAGG	GGTSLAAPLFTGFWA
Chromobacterium violaceum ATTC 12472	34498974	GQGATIGIVTLASFTPSDAFQ	SSETTLDVEQSGGIAPD	GGTSFVAPGLAGITA
Clostridium acetobutylicum ATTC 824D	15893913	GKNESIGIVTLAEFNPNDAYS	ADETTLDVEQSGALAPK	GGTSIVAPQLAGLCA
Erwinia carotova atroseptica SCRII043	50120389	GAGQCIGIIELGGGYRLPQLE	IDEVQMDIEIAGTLAPA	GGTSAVAPLWAGLLA
Leifsonia xyli subsp. *xyli* CTCB07	50954460	GAGTKVAIVAAFDDPAVAANT	TEEQHLDVQAVHAMAPD	GGDSLATPMVASMVA
Picrophilus torridus DSM_9790	48477259	GNGTTIVIVDAYGDPSINYDV	ATETALDVEWAHAIAPG	GGTSVATPIWAGIIA
Picrophilus torridus DSM_9790	48477281	GSGQSIGILDFYGDPFIKEEL	AGEISLDVESSHTMAPG	GGTSEASPILAGLMT
Picrophilus torridus DSM_9790	48478122	GOGITVAVIEVGDLPMSWLQE	TLETALDIEYIAAMAPD	GGTSFATPISAGEWA
Ralstonia eutropha JMP 134	73541448	GADRTIAIAEFGQNIGNGOVL	TAETMMDIEIVAGLCPK	GGTSAAAPLWAALVA
Streptomyces avermitilis MA-4680	29832492	GKGVTVAITDAYASPTIASDA	YGEETLDVEAVHAVAPK	GGTSLAAPVIAGVQA
Sulfolobus tokodaii 7	15921996	GEGYTIGILDFYGDPTIVQQL	NLEISLDVEVSHAMAPK	GGTSEASPLFAGLLT
Sulfolobus tokodaii 7	15922494	GSGVNIGILDFEGDPYIYQQL	ALEISLDVEYAHAAAPD	GGTSLATPIVAGIIA
Sulfolobus tokodaii 7	15922696	GNGTTVAIIDAYGDPTIYEDL	DIETALDVETVHAIAPY	GGTSLATPIVAGIIA
Sulfolobus tokodaii 7	15922823	GQNYTIGILDFYGDPYIAQQL	AGEISLDVEIAHTMAPE	GGTSEASPLTAGALV
Sulfolobus tokodaii 7	15922948	GKGSDIAIEGVPECYVNVSDI	SAENELDAEWSGAFSPG	YGTSGAAPMTAAMVS
Symbiobacterium thermophilum IAM 14863	51893408	GYGQTIGIIGIYHDYAEDAKA	YYEMALDVQAAHKMAPG	GATSVAAPMIAGVIA
Thermoplasma acidophilum DSM 1728	16081505	GQGITVAVIEVGFPIPSDMAQ	TLETSLDIEYIAAMAPM	GGTSFATPITAAEWA
Thermoplasma acidophilum DSM 1728	16082015	GMGETIGIVDAFGDPYLNYDI	IEETSLDVEWAHASAPY	GGTSLASPLWAGIIA
Thermoplasma acidophilum DSM 1728	16082551	GKGVKIGILGVGESANMSAIS	GVEADLDVEWSGAMAPN	GGTSFATPISAGIFA
Thermoplasma volcanium GSSI	13540979	GAGIKIGILGVGESANISAID	GVEADLDVEWSGAMAPN	GGTSFATPISAGIFA
Thermoplasma volcanium GSSI	13541541	GQGITVAVIEVGFPIPSDMAQ	TLETELDIEYIAAMAPM	GGTSFATPITAGEWA
Thermoplasma volcanium GSSI	13541953	GKGETIAIIDAYGDPFLNYDL	AQETSLDVEWAHVSAPL	GGTSLAAPLWAGVIA
Thermoplasma volcanium GSSI	13542325	GTGTTIAIVMGTGYINATLAY	LGEFTLDTQYSATVAPD	GGTSEATPTTAGMIA
Xanthomonas axonopodis citri 306	21241323	GSGSTIAILIGSDVLDSDIAT	AREATLDVDMALGGAPG	IGTSVAAPEFASVAA
Xanthomonas axonopodis citri 306	21241565	ASDIVGAVIAAGNLEQTLVRL	EVEWDMDTQLLVGSAGG	WGTSAATPTFAGYIA
Xanthomonas axonopodis citri 306	77748661	GHGQCIGIIVLGGGYARDQMT	DVEAQMDIQIAGAIAPG	GGTSAAAPLWAALLA
Xanthomonas oryzae KACC10331	77760762	GHGQCIGIIVLGGGYAREQMA	DVEAQMDIQIAGALVPG	GGTSAAAPLWAALLA
Other NCBI hits				
Acidothermus cellulolyticus 11B	88931817	GAGVTVALPEFEPFLSSDIAA	SGEAALDIETVAALAPS	GGTSAAAPLWAALLA
Acidothermus cellulolyticus 11B	88932005	GTGITVGITDAYASPTIAADA	FGEETLDVEAVHAMAQG	GGTSLAAPLFAGMTA
Alicyclobacillus sendaiensis	25900987	GQGQCIAIIELGGGYDEASLA	DGEVELDIEVAGALAPG	GGTSAVAPLFAALVA
Ferroplasma acidarmanus Ferl	68140013	GNNTTIVIVDAYGDPTLNYDV	ASETAIDVEWAHAIAPG	GGTSISTPMWAGIIA
Methylocapsa acidiphila	83308754	YGSAVIAIVDAYHNSSALADL	AGEEALDLDMAHALAPN	GGTSVSSPLVAALTN

TABLE II. (Continued)

Organism	Accession (GI code)	"Normal Asp-region"	Glu-Asp-region	Ser-region
Ralstonia solanacearum UM551	83749561	GAGQTIYIV**D**AMSDPNAAAEL	AT**E**IAL**D**VQWAHATAPL	GGT**S**LATPQWAGLLA
Rhodoferax ferrireducens DSM 15236	74023582	GAGQTIYIV**DD**AYNHPNVVKDL	AE**E**IAL**D**TQWAHAIAPL	GGT**S**LATPMWAAAVT
Solibacter usitatus Ellin6076	67865922	GAGQTVAII**E**LGGGYRTADLN	DG**E**VLL**D**IEVVGAIAPG	GGT**S**AVAPLWAALIA
Solibacter usitatus Ellin6076	67933815	GTGQLOAOV**G**ESDIDLSDIRA	WF**E**AD**L**DIEWAGAIARG	GGT**S**ASTPAFAGIVA
Solibacter usitatus Ellin6076	67927822	GTGQKIAIA**G**EWNLNLTDVRS	LL**E**AD**L**DVEYAGAVARN	GGT**S**AAAPAFAGIVA
Solibacter usitatus Ellin6076	67931923	GTGQKIAIA**G**QTQVDVADIQK	LG**E**AD**L**DLEWAGAVAPQ	GGT**S**AGTPAFAGITA

Conserved regions around catalytic residues.
*Other species and strains of Burkholderia, Sulfolobus and Xanthomonas have very similar sequence.

that this set represents a novel subfamily with different conserved residues than the D-H-S or ED-S families. They are found in phylogenetically diverse organisms (Table III). As yet, *Methanospirillum hungatei* is the only prokaryote predicted to have exclusively members of the E-D-S subfamily. All members of this new family were found iteratively using new HMMs for the regions surrounding putative catalytic residues (see later).

When compared to members of the sedolisin family in a multiple alignment (Fig. 2), it is clear that residues equivalent to the catalytic Ser278 and Glu78 are invariable, but neither the Asp/Glu32 nor Asp82 are present. Instead, at position 82 a Met is highly conserved, and at position 32 and 33 a Ser-Asp pair, but in both cases they are not 100% conserved (Table III, Fig. 2). The oxyanion residue is a conserved Asn164, in contrast to Asp164 of the sedolisins.

A closer inspection of the sequence alignment of this subfamily revealed a novel invariable Asp residue at the position equivalent to Ser128 in kumamolisin (or Ser125 in subtilisin) (Fig. 2). Homology modeling of the active site [Fig. 1(d)] shows that an Asp at position 128 would be in a very favorable position to form hydrogen bonds with Ser278 and Glu78, thereby forming a new alternative for Asp82 in stabilization of the general base. In this scenario, Ser32 could be involved in a larger stabilizing network through hydrogen bonds with intermediate water molecules. It is even conceivable that Asp128 could serve as the general base, with Glu78 providing a hydrogen-bonded link to Ser32, although this would require different orientations of side chains compared to the model in Figure 1(d). The semi-conserved Asp33 of this subfamily is presumably not involved in the stabilizing network since the model predicts it is oriented away from the other network partners.

E-H-S Family

Another set of 18 subtilase sequences was found that scored well with the HMMs for the regions surrounding the catalytic His and Ser, but these had a Glu residue at position 32 instead of an Asp (Table IV). Possibly this Glu32 serves the same function as Asp32 and stabilizes the general base His as part of the catalytic triad, as modeled in Figure 1(b). This homology model was made from kumamolisin as template, since the carboxylate group of Glu32 in kumamolisin is in nearly the same topological position as that of Asp32 in subtilisin.[32,33] Comparison of the template structures of subtilisin and kumamolisin shows that the backbone β-sheet strands are superimposable up to residue 31, but then the following loops deviate and differ in length by one residue, allowing the carboxylate side-chain groups to become topologically equivalent. In three cases this loop appears to be eight residues longer, before the residues HPDL topologically equivalent to the subtilisins are encountered (Table IV). *Nostoc* sp. gene gi:17227860 is a perfect example of a simple D32 to E32 substitution and an extra inserted residue in the following loop, with the sequence being otherwise highly similar to *Nostoc* sp. gene gi:17229107, which is a D-H-S family member:

TABLE III. The E-D-S Subgroup

Organism	Accession (GI code)	"Normal Asp-region"	Glu-Asp-region	New Asp-region	Ser-region
Consensus normal D-H-S subtilases		GkGvtVAVi**D**DtGvd-ynHpdL	xx**H**Gthvagiiag	NMDVINM**S**LGGPGTS	sGT**S**mAaPhvaGvaA
Genome hits					
Gloeobacter violaceus PCC 7421	37520846	GTGIKIGVL**S**DSYNCQGAAAA	SD**E**GRAMLQIVHDLAPG	GCTVIVD**D**VEYFNES	FGT**S**AAAPHAAAIAA
Gloeobacter violaceus PCC 7421	37522729	GSGITVGVL**S**DSYNTSTNPVK	ID**E**GRAMLQIIHDLAPK	GASVIVD**D**IIYLDEP	FGT**S**AAAPHAAAIAA
Gloeobacter violaceus PCC 7421	37522730	GGGITVGAL**S**DSYDTAAVDLG	ID**E**GRAMLQIIHDLAPK	GASVIVD**D**IIYLSEP	FGT**S**AAAPHAAAIAA
Methanosarcina acetivorans	20090865	GTGIKIGII**S**DGVDNLEDVQA	GN**E**GTNMLEIVYDIAPG	GCTVICD**D**IGWLAEP	YGT**S**ASCPHVAAIAA
Methanosarcina acetivorans	20093024	GAGIKIGII**S**AGVEDISEAIN	GN**E**GIVMLEIVHETSPG	GCQILCD**D**VGWPDEP	TGT**S**ASAPSVAGIGA
Methanosarcina mazei	21227078	GTGIKIGII**S**DGVEDISEADR	GT**E**GTVILEVVHKVSPG	GCQIICD**D**VGWPDEP	AGT**S**ASAPSVAGIGA
Methanospirillum hungatei JF - 1	88603735	GKGIKVGVI**G**NGAESLELSQK	GD**E**GTAMLEIIHDIAPD	GCRIICD**D**LYFFKQP	PGT**S**AAAPHVAGVIA
Methanospirillum hungatei JF - 1	88602240	GAGVIVGVV**S**SGVKGLADAQR	KA**E**GTAMMEIIHDIAPG	GATIIVE**D**VFNYEVP	TGT**S**AAAPHIAGLLA
Methanospirillum hungatei JF - 1	88602350	GEGVKVGVI**S**DGVDGLEDLKA	GD**E**GLAMLQIIHDIAPN	GCNIICD**D**ITYV-EP	TGT**S**AAAPHIAGLAA
Methanospirillum hungatei JF - 1	88602238	GSGIGIGII**S**NGAAGLIQAQE	GS**E**GTAMMEIIHDIAPG	GARIIVD**D**VGFLQVP	PGT**S**AAAPHIAGLLA
Ralstonia solanacearum GMI1000	17547820	GKGITVGLI**S**DSFNCNSQLNQ	TD**E**GRAMAEIIHDVAPG	GAQVIVD**D**LQYSYEP	YGT**S**AAAPHVAGVAA
Ralstonia solanacearum GMI1000	17548824	GKGITVGVL**S**DSFNCNSERNQ	GD**E**GRGMAEIIHDVAPG	GAQIIVD**D**VEYFEEP	LGT**S**AAAPHLAAVAA
Rhodobacterpirellula baltica SPI	32476420	GAGIKIGVI**S**DSYSRTNGGGG	KD**E**GRAMLELIHDIAPG	GVDIIVD**D**VTYAGMQ	AGT**S**AAAPNAAAVAA
Salinibacter ruber DSM13855	83814483	GSGQKICAL**S**DSYDARGQASR	SD**E**GRAMLQLIHDIAPG	GCTVIVD**D**VGYNLEP	FGT**S**AAAPNVAAIGA
Other hits					
Blastopirellula marina DSM 3645	87285449	GTGQKIGVI**S**DTYNADGSALL	TD**E**GRAMLQLVHDIAPG	GSTVIVD**D**IGFSNEP	FGT**S**AAAPHVAALAA
Bradyrhizobium sp. BT Ail	78692768	GKGIKIGLL**S**DSFDFLKGADA	ID**E**GRAMLQIVHTIAPG	GCKVICD**D**IFYYHEP	YGT**S**AATPTVAALAA
Janibacter sp. HTCC2649	84498360	GSGIDVGVI**S**DGVTSIAAAQA	GD**E**GTSMLEIVHDMAPG	GVDIITE**D**IPFDSEP	FGT**S**AATPSSAGVAA
Ralstonia solanacearum UW551	83746410	GKGITIGVI**S**DSFNCNSELNQ	TD**E**GRAIAEILHDVAPG	GAQVIVD**D**MQYSYEP	YGT**S**AAAPHLAGVAA

Conserved regions around catalytic residues.

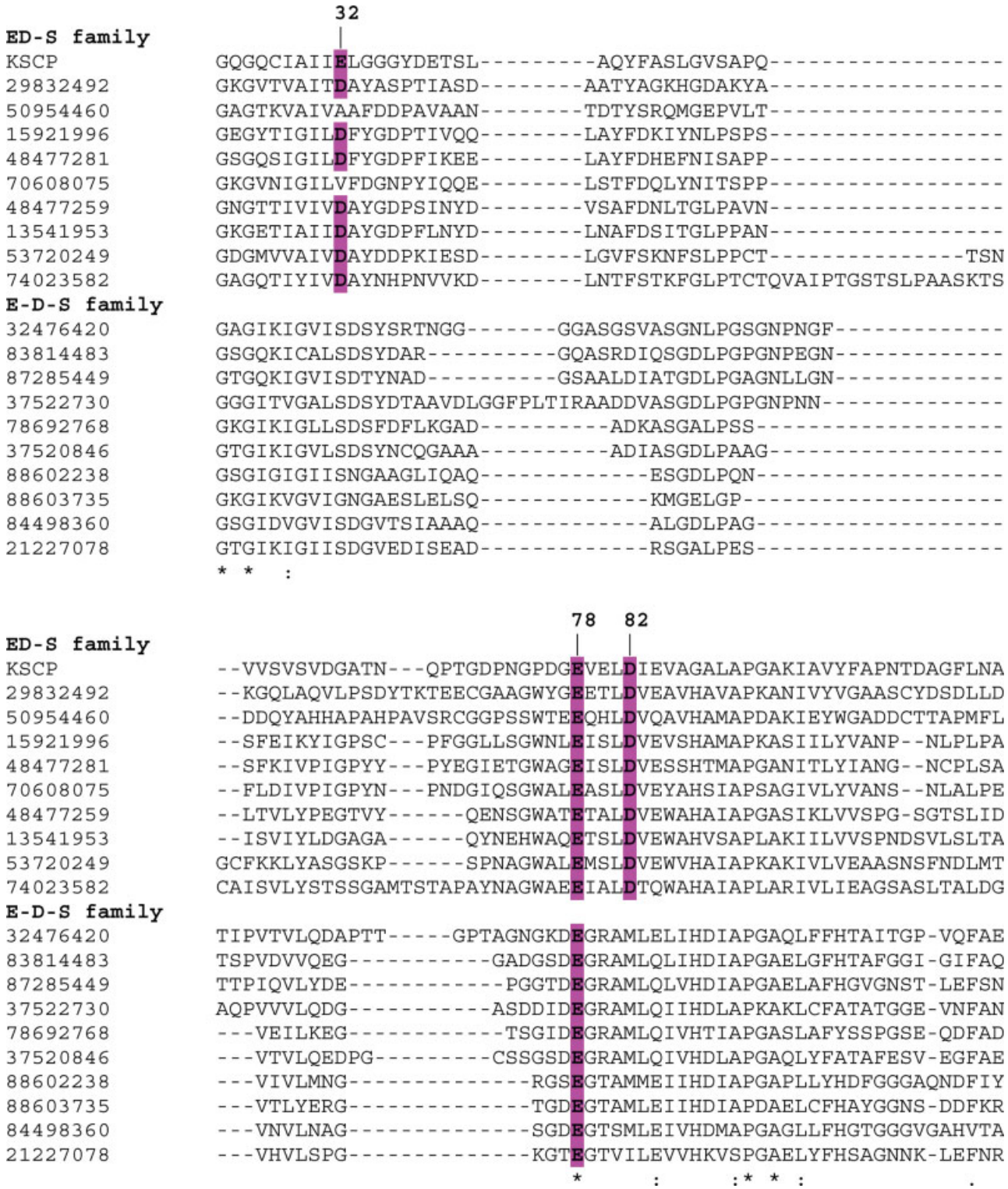

Fig. 2. Multiple (trimmed) sequence alignment and comparison of selected members of the ED-S and E-D-S families. NCBI codes of proteins are shown. KSCP represents the sequence of kumamolisin from *Bacillus* novosp. MN-32 for which the crystal structure has been determined.[32,33] Positions and numbering of the kumamolisin catalytic residues Glu32, Glu78, Asp82, and Asp164 are shown. The proposed new catalytic Asp residue in the E-D-S family corresponds to residue Ser128 in the ED-S family (sedolisins). Putative catalytic residues are shaded purple, while the putative oxyanion hole residues is shaded yellow.

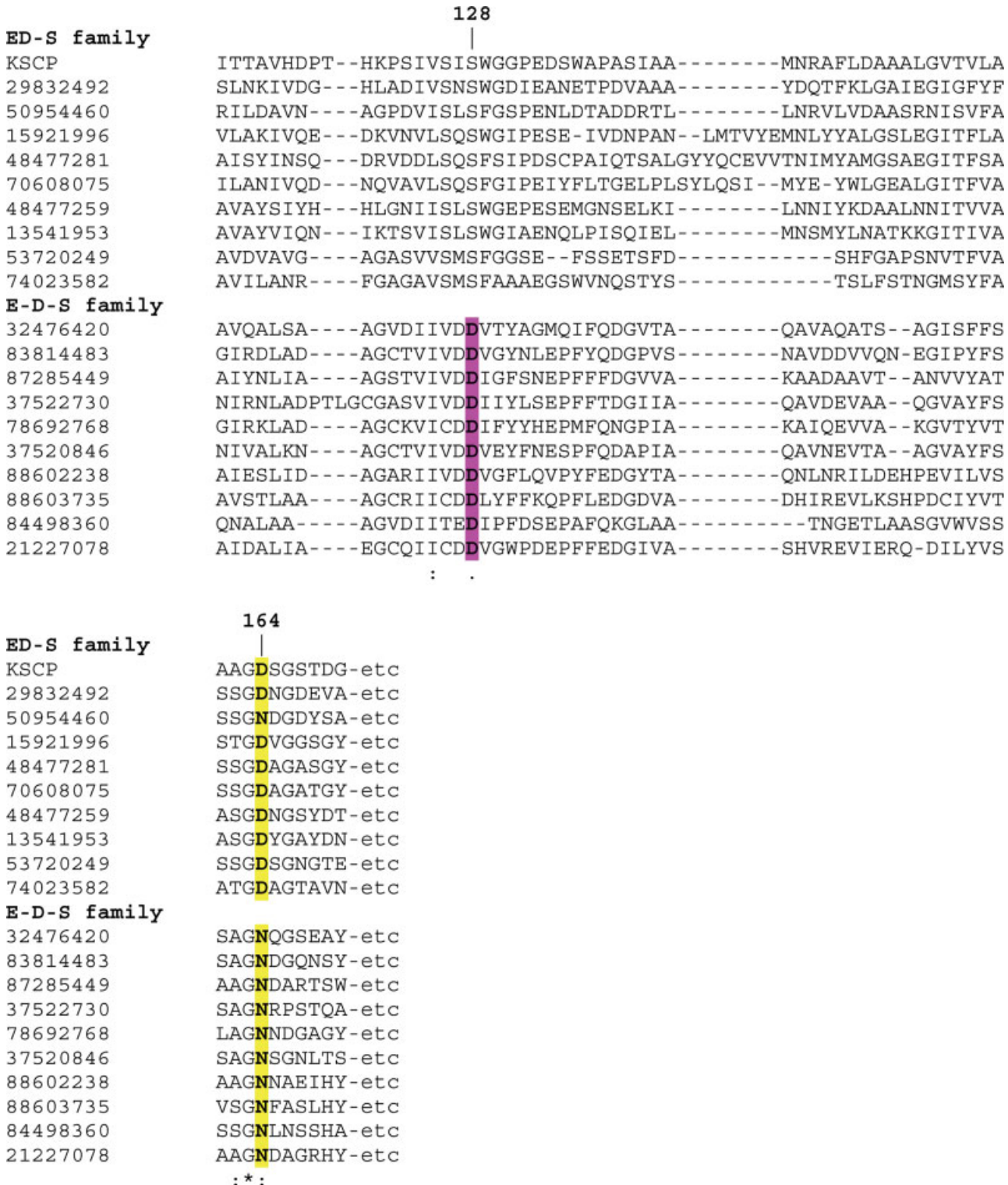

Figure 2. (Continued.)

TABLE IV. The E-H-S Subgroup(s)

Species	Accession	Glu region	His region	Ser region
Consensus normal D-H-S subtilases		GKGvtVAViDtG-vdynHpdL	HGthvagiiag	sGTSmAaPhvaGvaAlll
Genome hits				
Bacillus anthracis	45729180	GSGITFVDMEYG-WLLNHEDL	HGTSVLGIVSS	SGTSSASPIIAGAATLVQ
Bacillus cereus ATTCC14579	30021516	GQGATFVDLEEG-WLLNHEDL	HGTSVLGVVSA	RGTSSASPIIAGAAVSIQ
Bacillus cereus ATTCC14579	30021855	GNGITFVDMEYG-WLLNHEDL	HGTSVLGIVSS	SGTSSASPIIAGAATLVQ
Bacillus cereus ATCC10987	42782844	GSDVTFVDMEYG-WLLNHEDL	HGTSVLGIVSS	SGTSSASPIIAGAATLVQ
Streptomyces avermitilis	29833191	GQDVTVIDVEGA-WQLGHEDL	HGTAVIGVIGG	SGTSSASPMVVGALAALQ
Anabaena variabilis ATCC29413	75910870	GRKIAIGQVEIGRPGMFGWDK	HAYNVAGVMVS	TGTSFAAPHLTATVALLQ
Nostoc sp. PCC7120	17227860	GRGVTVGVFEGGGVEYTHPDL	HATSVAGVIGA	NGTSAAAPEVSGVVALML
Mycoplasma gallisepticum R	31544301	EKRIGVAVLEVGERENDSKAL	HSTKVGSIISG	SGTSFSAPFISGILANTL
Mycoplasma gallisepticum R	31544303/4	NKRVGVAVLEIGE-GFLQAQA	HATAVASIISG	YGTSFSAPFISGVIANTL
Mycoplasma gallisepticum R	31544314	QKRIGIAVLEVGE-GDKHPER	HSTEVASVISG	SGTSYSAPFVSGVLANTL
Mycoplasma gallisepticum R	31544366	EKRIGVAVLEVGESYDMRKAL	HATEVGSVISG	YGTSFASPFVSGVLANTL
Mycoplasma gallisepticum R	31544876/7	QERIGVAILEASN-REDRTKA	HATKVAAIVSG	QGTSFSAPFVSGVIANTL
Mycoplasma hyopneumoniae 232	54020128	SPQTKVGAIEVKH-EFNYNFM	HSTLVSLILGS	NGTSFAAPIVTGLISTLL
Mycoplasma hyopneumoniae 7448	72080669	SPQTKVGAIELWD-EFNYNFI	HSTLVSLILGG	NGTSFAAPVVTGLISTLL
Mycoplasma hyopneumoniae J	71893444	SPQTKVGAIEVTD-EFNYNFM	HSTLVSLILGG	SFTSFAAPVVTGLISTLL
Mycoplasma hyopneumoniae J	71893686	APRERVGVVEAD-MSGTFDEN	HATLVSGIIGG	SGTSFSAPIVTGIISTID
Photorhabdus luminescens TTOJ	37524644	GKGVRIGQFEPGGKFATAPEIFDINHPDL	HATMVAGVMVA	QGTSFAAPIVSGVVALML
Pseudomonas aeruginosa	15596439	TRPVRIGVIERD-VDFDAPDF	HGSTVAGILAA	CGTSYSTPMVAGTVAAML
Pseudomonas putida KT2440	26990807	VKPVRVGVIERE-VDFDAPGF	HGSHVAGILAA	CGTSYATPLVTATVATML
Pseudomonas putida KT2440	26991602	GKGVRIGQFEPGGEFAVAPEIFDIGHPDL	HATQVAGVMVG	QGTSFAAPIVSGVVALML
Other NCBI hits[*]				
Yersinia bercovieri ATCC 43970	77956358	GKGVRIGQFEPGGQFATGPMIFDINHPDL	HATMVAGVMVA	QGTSFAAPIVSAIAALML
Crocosphaera watsonii WH 8501	67925119	GRKIAIGQVEIGRPGIFGFDK	HAAMVATVMVS	SGTSFAAPHITASVALLQ
Mixed group (ED-S group)				
Bacilllus novo sp. MN32 (KSCP)3D	21730221	GQGQCIAIIELGGGYDETSLA		
*Bradyrhizobium japonicum usda*110	27375805	GAGQCIAIIELNDIDQKGHPT		
Burkholderia mallei ATCC 23344	53716275	GAGQCIAIVELGGGYRPAEIQ		
Burkholderia pseudomallei K96243	53722583	GAGQCIAIVELGGGYRPAEIQ		
Erwinia carotova atroseptica SCRII043	50120389	GAGQCIGIIELGGGYRLPQLE		
*Picrophilus torridus*DSM_9790	48478122	GQGITVAVIEVGDLPMSWLQE		
*Ralstonia eutropha*JMP134	73541448	GADRTIAIAEFGQNIGNGQVL		
Thermoplasma acidophilum DSM 1728	16081505	GQGITVAVIEVGFPIPSDMAQ		
Thermoplasma volcanium GSSI	13541541	GQGITVAVIEVGFPIPSDMAQ		

Conserved regions around catalytic residues.
*And other species and strains of Yersinia.

```
17229107  -YTGQGVIVAVVDSG-VDYTHPDL-
17227860  -YTGRGVTVGVFEGGGVEYTHPDL-
```

The subtilases with this Glu-His-Ser triad are highly diverse in sequence similarity and length, and do not represent one clear subfamily. This is also evident from the region surrounding Glu32, which is not very well conserved (Table IV). This probably reflects different evolutionary subsets, with variations in loop orientation starting from residue Glu32. A new HMM for the Glu region was made from the sequences in Table IV, including those sequences from the ED-S family which also have Glu32. This Glu-HMM was used to identify new members of the E-H-S family, although scores are sometimes low due to the large sequence diversity in this region. Although some subtilases in Table IV already scored reasonably well with the classical Asp-HMM, most score better with the new Glu-HMM. A small subset, listed at the top of Table IV, has both an Asp and a Glu residue in this region, making it difficult to decide on the correct sequence alignment and the correct catalytic residue. In the suggested alignment, preferred by the Asp-HMM, the Asp30 residue carboxylate is structurally also oriented towards the Glu32 carboxylate, so that both could contribute to the hydrogen-bond network.

Mycoplasma is the only prokaryote with a strong preference for E-H-S family members, e.g. *M. gallisepticum* has five E-H-S members and only one D-H-S member.

Loss of Catalytic Residues

In *C. difficile*, three adjacent subtilase genes are found, of which two are fused as in *C. acetobutylicum* and *C. tetani.* The catalytic His and Ser residues in two of the *C. difficile* subtilase domains are both substituted (His to Gln/Thr, and Ser to Ala/Gly), presumably inactivating them, at least as serine proteases. Since both residues are replaced in adjacent genes, this argues against sequencing errors.

Another example of simultaneous mutation of catalytic residues is found in subtilases from five different *Rhodopseudomonas palustris* strains. In each case, concomitant mutations are seen of the catalytic residues His (to Gln, Ser, or Arg) and Ser (to Asn or Thr), and of the oxyanion hole Asn (to Ser or Arg). Substitution of the catalytic Ser residue was rarely found in other genomes, as the only two other examples observed were a replacement by Asp in *Thermobifida fusca* gene gi:72160625, and by Gly in *Mycobacterium avium paratuberculosis* gene gi:41409885. It stands to reason that more extensively modified regions around (and including) the catalytic residues will not be identified by the HMMs used by us.

Multiple Subtilases

It is more common to have multiple subtilase-encoding genes than a single gene, as can be seen in the Prokaryote SubtilaseDB. Several genomes were found to encode 10 or more subtilases, i.e. *Deinococcus radiodurans* (10 genes), *Streptomyces coelicolor* (11 genes), *Xanthomonas campestris* (11 genes), *Xanthomonas citri* (14 genes), *Bdellovibrio bacteriovorus* (15 genes), and *Streptomyces avermitilis* (15 genes). There are also variations in the number of subtilases genes found in different strains of a species (see the SubtilaseDB).

In a few instances it has been reported that two or more subtilase-encoding genes occur adjacent to each other on the chromosome, possibly even in the same operon.[9,34] In our genome-wide analysis we now find sets of two or more adjacent subtilase genes in 18 different species (Table V). In nearly all cases, adjacent genes are highly similar to each other (an average sequence identity of 56%; much higher when only subtilase domains are compared), suggesting one or more gene duplication events during evolution. This high similarity still holds when one or two other unrelated genes separate the subtilase genes, suggesting that an insertion has occurred after duplication of the subtilase genes. The best example is in *Geobacter metallireducens* where a regulator gene separates two nearly identical subtilase genes (85% identity overall, 99% in subtilase domain).

Annotation and Predicted Properties of Subtilases

Our genome-wide analysis allows the first annotation as proteases, and more specifically as subtilases, of over 100 proteins in different genomes. Of the 567 subtilases identified by us, 95 are currently annotated in the NCBI database as hypothetical proteins, and another 18 proteins are annotated with either a general, an unrelated, or an incorrect function (see Supplementary Table S3). Current general and unrelated annotations such as "membrane protein," "autotransporter," "TPR-repeat protein," or "fibronectin type III domain protein" could be partially correct, since we find these to be large proteins with other domains attached to the subtilase domain. Moreover, the large majority of subtilases are annotated in the NCBI database simply as prote(in)ase, peptidase, or serine protease (see Supplementary Table S4), and their annotation can now be improved by adding the terms subtilase, subtilisin-like, or subtilisin family, and more specifically by adding the subfamilies as defined by us (as indicated in Supplementary Table S3).

About 65% of the subtilases have a predicted signal peptide by SignalP,[28,35,36] and hence should be translocated across the cell membrane and function extracellularly. There are presumably more subtilases with a signal peptide, since some signal peptides are difficult to identify, particularly when the start codon has been chosen incorrectly. Surprisingly, only 27 of the subtilases have a predicted LPxTG motif for anchoring to the peptidoglycan layer, and these are nearly all in streptococci. Hence the majority of subtilases are presumably translocated across the cell membrane, but only a limited number are predicted to be covalently attached to the cell surface.

TABLE V. Adjacent and Fused Subtllase Genes in Genomes

Species	Family	Number of genes	Genes (NCBI accession code)	Comments
Bacillus licheniformis ATCC 14580	D-H-S	2	52080132/33	52080132 is highly similar to N-terminal part of 52080133; addional >900 resdues in letter may be result of gene fusion
Bacillus licheniformis DSM 13	D-H-S	2	52785506/07	52785506 is highly similar to N-terminal part of 52785507; addional >900 resdues in letter may be result of gene fusion
Chromobacterium violaceum	ED-S	2[a]	34497420/23	Highly similar; 2 very small intermediate genes
Clostridium acetobutylicum	D-H-S	1	15896490	2 fused subtilase genes; both active
Clostridium tetani	D-H-S	1	28211939	2 fused subtilase genes; both active
Clostridium difficle	D-H-S	2	ERGO codes	RDF01780 has 2 fused subtilase genes, 2nd domain is inactive; RDF01781 is also inactive and most similar to C-terminal domain of RDF01780
Clostridium perfringens	D-H-S	2	18311094/95	Highly similar, but also to 18311543/44/45
	D-H-S	3	18311543/44/45	All highly similar, but also to 18311094/95
Geobacter metallireducens	D-H-S	2[a]	78193224/26	Intermediate gene 78193225 encodes a regulater; protease domains are nearly 100% identical
Gloeobacter violaceus	E-D-S	2	37522729/30	Highly similar, also outside protease domain
Idiomarina loihinsis L2TR	D-H-S	2	56459272/73	Not similar; genes are oriented convergently
Methanospirillum hungatei JF-1	E-D-S	2[a]	88602238/40	Highly similar in protease domain; intermediate gene 88602239(457 aa) encodes a hypothetical protein
Mycoplasma gallisepticum	E-H-S	1 + 1[a]	31544301/ (303-304)[b]	Highly similar; intermediate gene 31544302(491 aa) encodes a unique hypothetical protein
Nitrosospira multiformis ATCC 25196	D-H-S	2[a]	82703009/12	Highly similar; intermediate genes 82703010 and 82703011 encodes homologous hypothetical proteins (223 aa)
Pseudomonas fluorescens Pf-5	D-H-S	2	70730567/68	Fairly similar, other domain (autotransporter) is highly similar
Pseudomonas fluorescens Pf0-1	D-H-S	2	77458908/09	Fairly similar, other domain (autotransporter) is highly similar
Pseudomonas syringae	D-H-S	2	28868855/56	Fairly similar, other domain (autotransporter) is highly similar
Ralstonia solanacearum	D-H-S	2	17547372/73	Highly similar
Streptomyces avermitilis	D-H-S	2	29832993/94	Highly similar
Xanthomonas campestris ATCC33913	D-H-S	3	21230325/26/28	Highly similar
Xanthomonas campestris 8004	D-H-S	2	66769679/81	Highly similar
Xanthomonas campestris vesicatoria 85-10	D-H-S	3	78046515/16/17	Highly similar, also to genes 78049225/27
	D-H-S	2[a]	78049225/27	Highly similar, also to genes 78046515/16/17; intermediate gene 78049226 encodes a hypothetical protein (2357 aa)

TABLE V. (Continued)

Species	Family	Number of genes	Genes (NCBI accession code)	Comments
Xanthomonas citri	D-H-S	3	21241698/699/700	Highly similar
	D-H-S	1 + 1[a]	21243558/60	Highly similar; intermediate gene 21243559 (202 aa) encodes a hypothetical protein
	D-H-S	1 + 1[a]	21244270/72	Highly similar; intermediate gene 21244271 (2190 aa) encodes a hypothetical protein

[a]There is a non-subtilase gene between 2 subtilase genes.
[b]Split gene.

DISCUSSION

The extreme sequence variability of subtilases has now been found to extend to two of their three catalytic triad residues. A genome-wide search for subtilases, with iteratively improved HMMs for regions surrounding catalytic residues, has led to the identification of at least four families with variations in catalytic residues. The nucleophile Ser is invariably found in all subtilases, while the nature *and* position (in the protein sequence) of the general base and acid residues of the catalytic triad are found in different combinations. Additional side chains may contribute to a stabilizing hydrogen-bond network, presumably increasing the potential of variations in catalysis and stability within this serine protease superfamily.

With the exception of the sedolisin family, such variations in the catalytic residues have not been described before in subtilases. This phenomenon has been described in other enzyme families, however. Variations in the catalytic triad residues in the α/β-hydrolase family are common, and lead to differences in catalytic mechanism and type of cleaved bonds.[37] The α/β-hydrolase fold provides a scaffold for the active sites of various enzymes, including proteases, lipases, esterases, dehalogenases, peroxidases, and epoxide hydrolases. The catalytic triad always consists of a highly conserved nucleophile (Ser, Asp, or Cys), an acidic residue (Asp or Glu), and a fully conserved His residue. Variations in the topological position of the acidic residue have also been found in α/β-hydrolases.[37]

Based on our present observations, we propose that subtilases have also evolved this flexibility in catalytic residues, both in type and their topological position. The simplest adaptation appears to be the replacement of Asp32 by Glu, as we have found in the E-H-S family members (Table IV). The high variability in the residues surrounding Glu32 suggests some fold variability in this region as well, possibly leading to differences in specificity, since residue 32 is located in the P2-binding pocket of subtilases.[2,11] More drastic is the replacement of the catalytic His by Glu, combined with a topologically different Asp residue than at position 32. We propose that two different scenarios have evolved for the position of this stabilizing Asp residue. The first case is the structurally characterized sedolisin family (ED-S family), in which the Asp is four residues downstream of His, positioned on the same helix (i.e. His78 and Asp 82 in kumamolisin). Together with an Asn to Asp substitution in the oxyanion hole, this leads to enzymes of acidic pH optimum, both endopeptidases and tripeptidylpeptidases, as determined experimentally.[18,30,31] In the second scenario, the E-D-S family first described in this work, the stabilizing Asp is predicted to be at the end of a different β-strand, in a position topologically equivalent to Ser125 of subtilisin. The oxyanion hole residue is still Asn in this subset of subtilases. Although there is no experimental evidence as yet to support this hypothesis, our homology modeling indicates that an Asp at this position could be favorably oriented to contribute to a stabilizing proton-transfer network. These substitutions of catalytic residues have a wide phylogenetic distribution, suggesting that they are not species or branch-specific.

Simultaneous loss of the catalytic residues His and Ser was found in duplicated and fused genes in *Clostridium* and *Rhodopseudomonas*. This could reflect an evolutionary process ultimately leading to enzymes with different catalytic mechanisms and specificities or even nonenzymatic functions. When the latter stage of sequence variability has been reached, the identification of distant family members based on sequence motif conservation, such as with our HMMs, becomes very fuzzy and should be replaced by structural-fold comparison search methods.

It should be particularly interesting to determine experimentally whether these subtilases with variations in active site residues are still functionally active as proteases, or whether they have evolved to new enzymatic or other functions as in the α/β-hydrolases.

The proposed new division into subtilase families, their HMMs, and identified gene sets will be communicated to various databases such as Merops, PROSITE, Pfam, etc.

ACKNOWLEDGMENTS

We thank Quinta Helmer and Klaas Schotanus for genome database searches.

REFERENCES

1. Rawlings ND, Morton FR, Barrett AJ. MEROPS: the peptidase database. Nucleic Acids Res 2006;34:D270–D272.

2. Siezen RJ, Leunissen JA. Subtilases: the superfamily of subtilisin-like serine proteases. Protein Sci 1997;6:501–523.
3. Bergeron F, Leduc R, Day R. Subtilase-like pro-protein convertases: from molecular specificity to therapeutic applications. J Mol Endocrinol 2000;24:1–22.
4. Beers EP, Jones AM, Dickerman AW. The S8 serine, C1A cysteine and A1 aspartic protease families in Arabidopsis. Phytochemistry 2004;65:43–58.
5. Antao CM, Malcata FX. Plant serine proteases: biochemical, physiological and molecular features. Plant Physiol Biochem 2005; 43:637–650.
6. Cheng Q, Stafslien D, Purushothaman SS, Cleary P. The group B streptococcal C5a peptidase is both a specific protease and an invasin. Infect Immun 2002;70:2408–2413.
7. Siezen RJ, Kuipers OP, de Vos WM. Comparison of lantibiotic gene clusters and encoded proteins. Antonie Van Leeuwenhoek 1996;69:171–184.
8. Coutte L, Antoine R, Drobecq H, Locht C, Jacob-Dubuisson F. Subtilisin-like autotransporter serves as maturation protease in a bacterial secretion pathway. EMBO J 2001;20:5040–5048.
9. Shimamoto S, Moriyama R, Sugimoto K, Miyata S, Makino S. Partial characterization of an enzyme fraction with protease activity which converts the spore peptidoglycan hydrolase (SleC) precursor to an active enzyme during germination of *Clostridium perfringens* S40 spores and analysis of a gene cluster involved in the activity. J Bacteriol 2001;183:3742–3751.
10. Gey Van Pittius NC, Gamieldien J, Hide W, Brown GD, Siezen RJ, Beyers AD. The ESAT-6 gene cluster of *Mycobacterium tuberculosis* and other high G+C gram-positive bacteria. Genome Biol 2001;2:RESEARCH0044.
11. Siezen RJ, de Vos WM, Leunissen JA, Dijkstra BW. Homology modelling and protein engineering strategy of subtilases, the family of subtilisin-like serine proteinases. Protein Eng 1991;4: 719–737.
12. Siezen RJ. Multi-domain, cell-envelope proteinases of lactic acid bacteria. Antonie Van Leeuwenhoek 1999;76:139–155.
13. Boekhorst J, de Been MW, Kleerebezem M, Siezen RJ. Genome-wide detection and analysis of cell wall-bound proteins with LPxTG-like sorting motifs. J Bacteriol 2005;187:4928–4934.
14. Schneewind O, Mihaylova-Petkov D, Model P. Cell wall sorting signals in surface proteins of gram-positive bacteria. EMBO J 1993;12:4803–4811.
15. Janulczyk R, Rasmussen M. Improved pattern for genome-based screening identifies novel cell wall-attached proteins in gram-positive bacteria. Infect Immun 2001;69:4019–4026.
16. Wlodawer A, Li M, Gustchina A, Tsuruoka N, Ashida M, Minakata H, Oyama H, Oda K, Nishino T, Nakayama T. Crystallographic and biochemical investigations of kumamolisin-As, a serine-carboxyl peptidase with collagenase activity. J Biol Chem 2004;279:21500–21510.
17. Wlodawer A, Li M, Dauter Z, Gustchina A, Uchida K, Oyama H, Dunn BM, Oda K. Carboxyl proteinase from *Pseudomonas* defines a novel family of subtilisin-like enzymes. Nat Struct Biol 2001;8:442–446.
18. Reichard U, Lechenne B, Asif AR, Streit F, Grouzmann E, Jousson O, Monod M. Sedolisins, a new class of secreted proteases from *Aspergillus fumigatus* with endoprotease or tripeptidyl-peptidase activity at acidic pHs. Appl Environ Microbiol 2006; 72:1739–1748.
19. Wlodawer A, Li M, Gustchina A, Oyama H, Dunn BM, Oda K. Structural and enzymatic properties of the sedolisin family of serine-carboxyl peptidases. Acta Biochim Pol 2003;50:81–102.
20. Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O. The Comprehensive Microbial Resource. Nucleic Acids Res 2001;29:123–125.
21. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR The Pfam protein families database. Nucleic Acids Res 2004;32:D138–D141.
22. Durbin R, Eddy S, Krogh A, Mitchison G. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge: Cambridge University Press; 1998.
23. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 2003;31:3497–3500.
24. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
25. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 2004;5:113.
26. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 2003;52:696–704.
27. von Heijne G. The signal peptide. J Membr Biol 1990;115:195–201.
28. Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: Signal P 3.0. J Mol Biol 2004;340: 783–795.
29. Guo H, Wlodawer A. A general acid-base mechanism for the stabilization of a tetrahedral adduct in a serine-carboxyl peptidase: a computational study. J Am Chem Soc 2005;127:15662–15663.
30. Oyama H, Hamada T, Ogasawara S, Uchida K, Murao S, Beyer BB, Dunn BM, Oda K. A CLN2-related and thermostable serine-carboxyl proteinase, kumamolysin: cloning, expression, and identification of catalytic serine residue. J Biochem (Tokyo) 2002; 131:757–765.
31. Oda K, Nakatani H, Dunn BM. Substrate specificity and kinetic properties of pepstatin-insensitive carboxyl proteinase from *Pseudomonas* sp. No 101. Biochim Biophys Acta 1992;1120:208–214.
32. Comellas-Bigler M, Fuentes-Prior P, Maskos K, Huber R, Oyama H, Uchida K, Dunn BM, Oda K, Bode W. The 1.4 Å crystal structure of kumamolysin: a thermostable serine-carboxyl-type proteinase. Structure 2002;10:865–876.
33. Comellas-Bigler M, Maskos K, Huber R, Oyama H, Oda K, Bode W. 1.2 Å crystal structure of the serine carboxyl proteinase pro-kumamolisin; structure of an intact pro-subtilase. Structure 2004; 12:1313–1323.
34. Schmidt BF, Woodhouse L, Adams RM, Ward T, Mainzer SE, Lad PJ. Alkalophilic *Bacillus* sp. strain LG12 has a series of serine protease genes. Appl Environ Microbiol 1995;61:4490–4493.
35. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Int J Neural Syst 1997;8:581–599.
36. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 1997;10:1–6.
37. Nardini M, Dijkstra BW. α/β Hydrolase fold enzymes: the family keeps growing. Curr Opin Struct Biol 1999;9:732–737.