Protein Science (1997), 6:501–523. Cambridge University Press. Printed in the USA.
Copyright © 1997 The Protein Society

REVIEW

Subtilases: The superfamily of subtilisin-like serine proteases

ROLAND J. SIEZEN[1] AND JACK A.M. LEUNISSEN[2]

[1]Department of Biophysical Chemistry, NIZO, P.O. Box 20, 6710BA Ede, The Netherlands
[2]CAOS/CAMM Center, University of Nijmegen, Toernooiveld, 6525ED, Nijmegen, The Netherlands

(RECEIVED August 22, 1996; ACCEPTED November 5, 1996)

Abstract

Subtilases are members of the clan (or superfamily) of subtilisin-like serine proteases. Over 200 subtilases are presently known, more than 170 of which with their complete amino acid sequence. In this update of our previous overview (Siezen RJ, de Vos WM, Leunissen JAM, Dijkstra BW, 1991, *Protein Eng* 4:719–731), details of more than 100 new subtilases discovered in the past five years are summarized, and amino acid sequences of their catalytic domains are compared in a multiple sequence alignment. Based on sequence homology, a subdivision into six families is proposed. Highly conserved residues of the catalytic domain are identified, as are large or unusual deletions and insertions. Predictions have been updated for Ca^{2+}-binding sites, disulfide bonds, and substrate specificity, based on both sequence alignment and three-dimensional homology modeling.

Keywords: homology modeling; sequence alignment; serine protease; subtilase; subtilisin family

Serine endo- and exo-peptidases are of extremely widespread occurrence and diverse function. Many distinct families of serine proteases exist; they have been grouped into six clans (Rawlings and Barrett, 1994; Barrett and Rawlings, 1995), of which the two largest are the (chymo)trypsin-like and subtilisin-like clans. These two clans are distinguished by a highly similar arrangement of catalytic His, Asp, and Ser residues in radically different β/β (chymotrypsin) and α/β (subtilisin) protein scaffolds.

In 1991, we presented a review of over 40 members of the subtilisin-like serine proteases, termed "subtilases," which occur in Archaea, Bacteria, fungi, yeasts, and higher eukaryotes (Siezen et al., 1991). The mature enzymes were found to contain up to 1775 residues, with N-terminal catalytic domains ranging from 268 to 511 residues, and signal and/or activation-peptides ranging from 27 to 280 residues. Several members contain C-terminal extensions, relative to the subtilisins, which display additional properties such as sequence repeats, Cys-rich domains, or transmembrane segments. From four known crystal structures and a multiple alignment of 40 known amino acid sequences, a core structure was predicted for the catalytic domain of all subtilases, together with the variations that are allowed in the main-chain length as a result of insertions and deletions (Fig. 1). Nineteen of these core residues were found to be highly conserved, 10 of which are glycines. Predictions were also made for subtilases of unknown three-dimensional structure concerning essential conserved residues, allowable substitutions, disulfide bonds, Ca^{2+}-binding sites, substrate-binding site residues, ionic and aromatic interactions, and surface loops. Based on these predictions, strategies for homology modeling and protein engineering were developed and implemented, aimed at modulating either stability, catalytic activity, or substrate specificity (Siezen et al., 1991, 1993, 1994, 1995a).

Since 1991, more than 100 new subtilases have been discovered, and these are now included in this updated review. In addition to many new enzymes from micro-organisms, numerous members of the subtilase superfamily have now also been identified in various eukaryotes such as slime molds, plants, insects, nematodes, molluscs, amphibia, fish, mammals, and even in a catfish virus.

Reprint requests to: Dr. Roland J. Siezen, Department of Biophysical Chemistry, NIZO, P.O. Box 20, 6710BA Ede, The Netherlands; e-mail: siezen@nizo.nl.

Structure-based alignment

The coordinates of subtilisin BPN', subtilisin Carlsberg, thermitase, and proteinase K were used previously (Siezen et al., 1991) to determine the core of "structurally conserved regions" (scrs; Greer, 1990) and the common secondary structure elements, as analyzed with the DSSP program (Kabsch and Sander, 1983). This core of about 190 residues contains virtually all of the common α-helix and β-strand elements, including the active site residues D32, H64, and S221 (Siezen et al., 1991). Slight adjustments to these core regions have now been incorporated (core ABC in Fig. 2) based on a recent spatial superpositioning of seven structures that also included mesentericopeptidase, Savinase, and Esperase (Heringa et al., 1995); topologically equivalent residues were defined as those that have Cα-atom distances of less than 2.0 Å. The "variable regions" (or vrs) nearly always correspond to

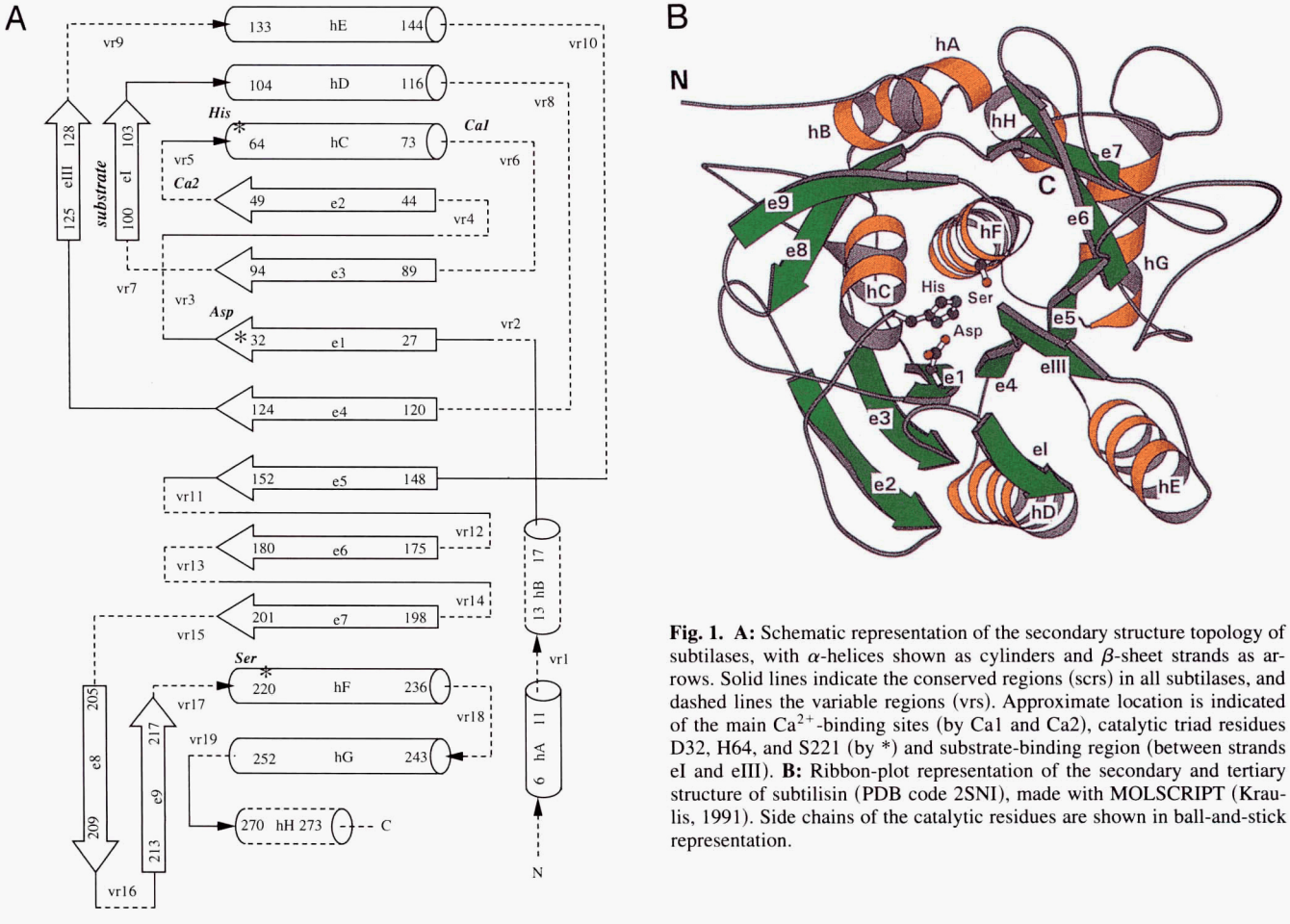

Fig. 1. A: Schematic representation of the secondary structure topology of subtilases, with α-helices shown as cylinders and β-sheet strands as arrows. Solid lines indicate the conserved regions (scrs) in all subtilases, and dashed lines the variable regions (vrs). Approximate location is indicated of the main Ca^{2+}-binding sites (by Ca1 and Ca2), catalytic triad residues D32, H64, and S221 (by *) and substrate-binding region (between strands eI and eIII). **B:** Ribbon-plot representation of the secondary and tertiary structure of subtilisin (PDB code 2SNI), made with MOLSCRIPT (Kraulis, 1991). Side chains of the catalytic residues are shown in ball-and-stick representation.

connecting loops between helices and strands and generally lie on the external surface of the protein (Fig. 1).

When only the subtilisin BPN′, subtilisin Carlsberg, and thermitase structures were superimposed the number of structurally equivalent Cα atoms increased to over 230 (or about 85% of all Cα atoms), which we refer to as the "extended core" (core AB in Fig. 2). This distinction between core and extended core scrs is of relevance for homology modeling, because the superfamily of subtilases can be subdivided into several families (see below).

Identification of subtilase superfamily members

An extensive search of scientific literature and databases (EMBL, Genbank, Swiss-Prot) was performed to identify new subtilisin-like serine proteases, using the programs BLAST (Altschul et al., 1990), TFASTA, and FASTA (Pearson and Lipman, 1988). Consensus sequence segments of 20–40 residues around the active site residues D32, H64, and S221 were used for this purpose; different consensus segments were obtained for different subtilase families (see Fig. 2). Sequences from patent literature and databases are not included because they represent synthetic or mutated genes encoding engineered subtilases. The main results of these searches are summarized in Tables 1 and 2. Further details, including reference to 10 crystal structures, can be found in the EMBL/Genbank and PDB databases using codes listed in the tables.

At present, over 170 complete and several partial amino acid sequences of subtilases are known; most are derived from the corresponding gene or cDNA sequences. We caution that in many cases it has not been established whether these genes encode functional proteins or whether the encoded protein is actually a protease. Examples of the latter are the outer-membrane antigen phssa1 of *Pasteurella haemolytica* (Lo et al., 1991), and the anti-freeze protein af70 of *Picea abies* (EMBL D86598), which were not described as proteases by the authors.

The majority of the subtilases are synthesized as pre-pro-enzymes, subsequently translocated over a cell membrane via the pre-peptide (or signal peptide), and finally activated by cleavage of the pro-peptide. A detailed comparison of the pre-pro sequences and the putative processing sites of these subtilases has identified two main types of pro-peptide (Siezen et al., 1995b). However, there are numerous exceptions in which the pro-peptides appear to be completely unrelated or even absent. A small number of subtilases is intracellular (Table 1).

Table 1 shows that the (putative) mature enzymes range in size from 266 to 1775 residues. The catalytic domain or module is defined as the segment with sequence homology to subtilisins; it is always located at the N-terminal end of the amino acid sequence directly after the pre-pro region. This review is focussed only on the catalytic domains.

Alignment of primary sequences

The multiple sequence alignment of the catalytic domains of over 120 subtilases is shown in Figure 2. Additional variants with <10%

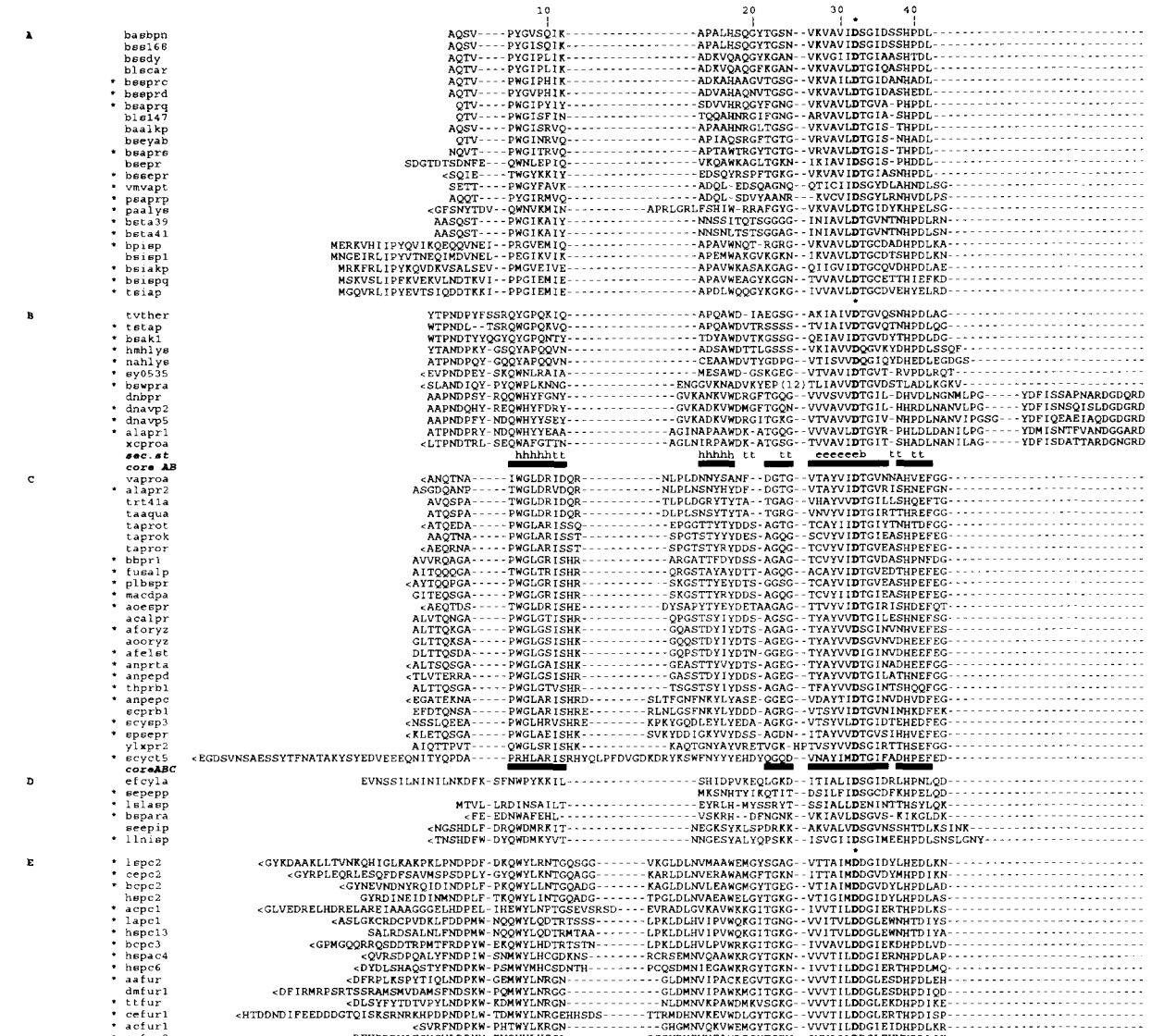

Fig. 2. *Continues on following pages.*

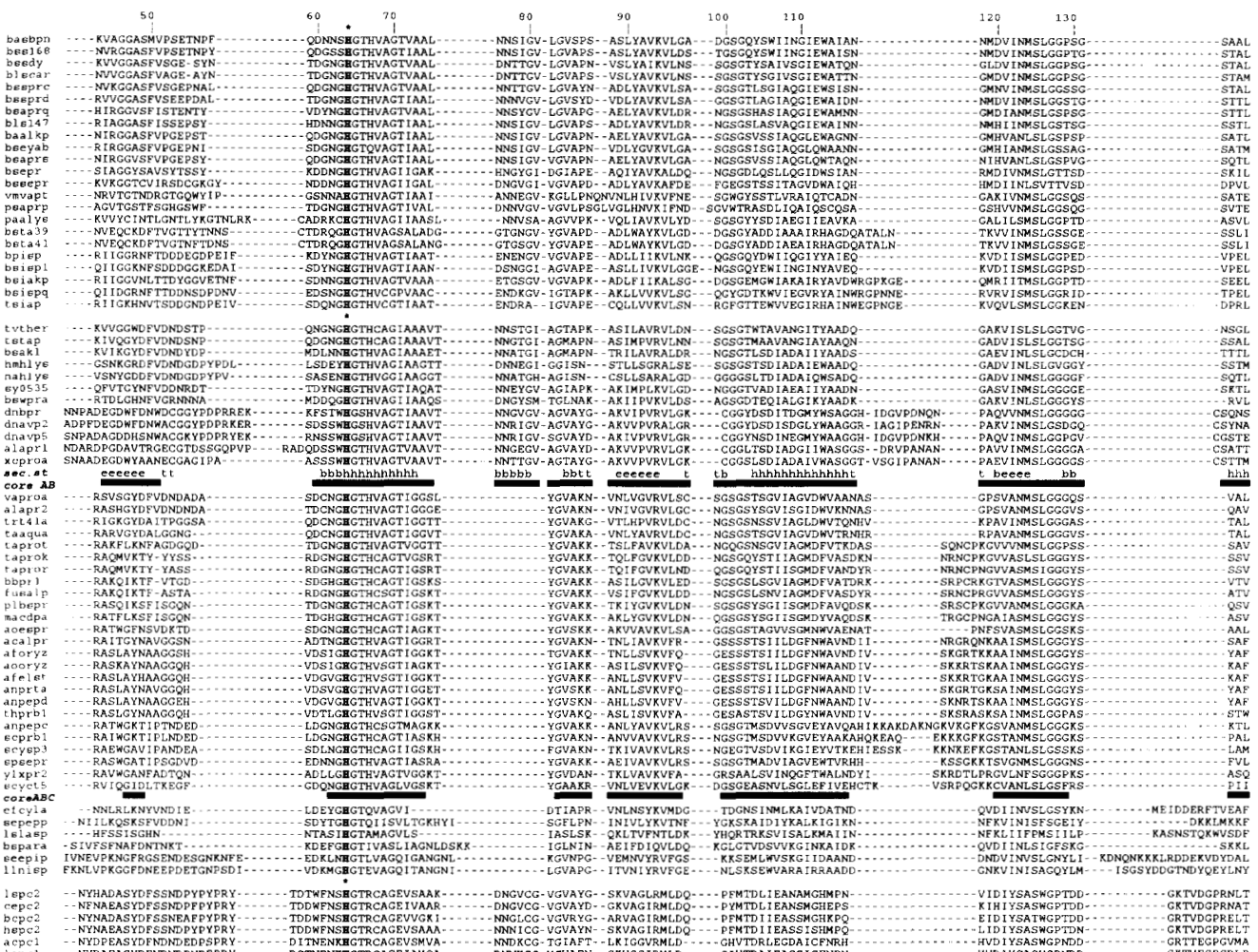

Fig. 2. *Continues.*

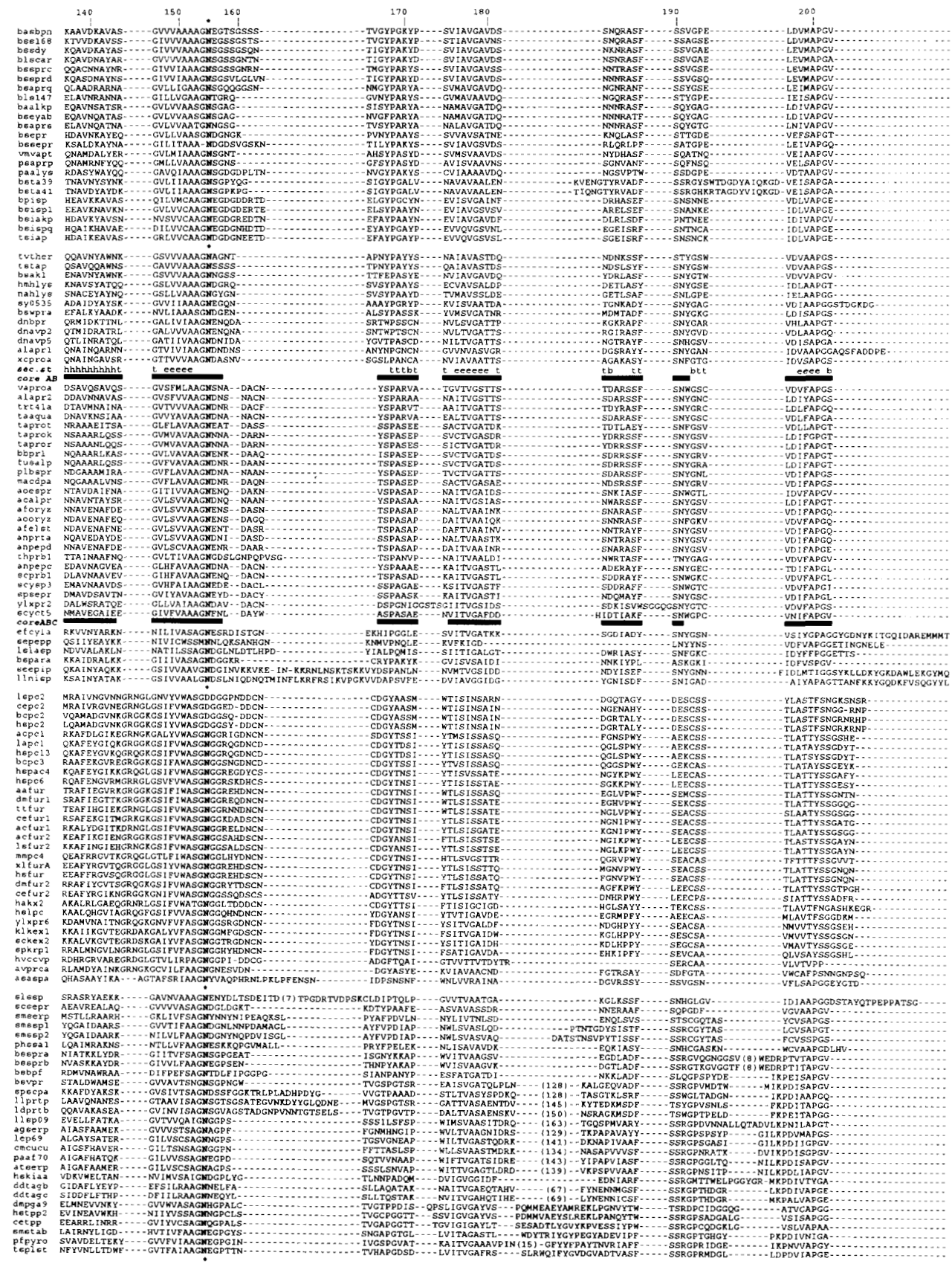

Fig. 2. *Continues.*

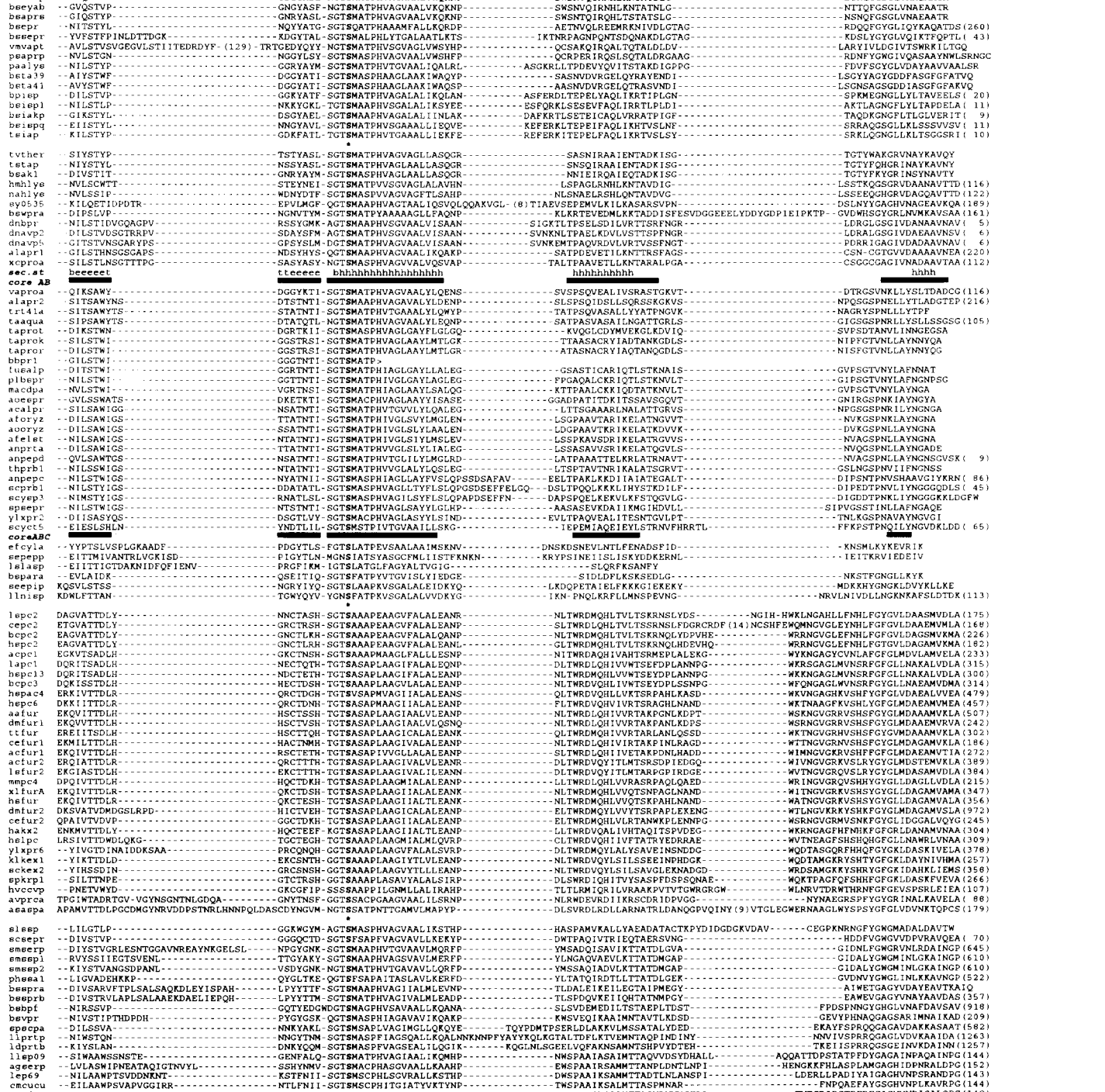

Fig. 2. *Continued (see facing page for caption).*

sequence difference are not shown in Figure 2 but are listed in Table 2. Amino acid numbering used throughout this review corresponds to that of mature subtilisin BPN′ (acronym basbpn), our reference sequence. Residues in inserts relative to this reference sequence are numbered in square brackets; for instance, residues inserted between positions 12 and 13 are numbered 12[+1], 12[+2], etc., or 13[−2], 13[−1] if more appropriate.

The conserved catalytic residues Asp 32, His 64, and Ser 221 are highlighted in Figure 2, as is the oxyanion-hole residue Asn 155. Conserved core elements (black bars) and secondary structure are indicated (Siezen et al., 1991, Heringa et al., 1995). This structural framework can be used for homology modeling of subtilases of known primary structures but unknown three-dimensional structures.

In some of the most highly diverged sequences there are regions with very weak sequence homology, even in the core, which results in alignments that are not unambiguous. In those cases, alternative alignments to those in Figure 2 may need to be considered. These regions are found on the surface of the molecule and contain numerous solvent-exposed residues, allowing for greater side-chain variation. Examples are (a) the exposed regions 43–58 and 182–218, which contain structurally conserved β-strands and turns; and (b) the exposed amphipathic helices 104–116, 133–144. and 243–252. In the latter case, the sequence alignment of amphipathic helices is also based on the requirement that at certain positions non-polar side chains are conserved that point into the interior of the molecule, while polar residues face outward. When necessary, correct three-dimensional positioning of Cys residues to form putative disulfide bonds was used as an aid in proper sequence alignment.

Sequence homology and family division

In Figure 3, the pairwise sequence identity within the catalytic domains is plotted graphically for all members of the subtilase superfamily aligned in Figure 2. It is clear that clustering occurs into groups or families, in which members show higher sequence identity to each other.

Figure 4 shows the parts of a family tree or cladogram, a measure of the sequence homology between superfamily members, constructed from the sequence alignment of the catalytic domains in Figure 2. In our earlier paper, a less extensive tree identified two main classes and some subclasses (Siezen et al., 1991). This expanded sequence information now allows a new subdivision into six families, which are summarized below. The dendrograms in Figure 4B illustrate the sequence homology within these families and further subdivision into subgroups (or subfamilies). Many of these subgroups are also apparent from the color patterns of sequence identity in Figure 3.

Subtilisin family

Only found in micro-organisms as yet. Includes mainly enzymes from *Bacillus*, with subgroups of true subtilisins (>64% identity), high-alkaline proteases (>55% identity), and intracellular proteases (>37% identity). Numerous minor variants of true subtilisins and high-alkaline proteases have been identified (Table 2). Long C-terminal extensions are rare. Several 3D structures are known (see Tables 1 and 2).

Thermitase family

Enzymes found only in micro-organisms, including some thermophiles (>55% identity) and halophiles. The characteristic N-terminal sequence was also found in several other *Bacillus* proteases (Table 3). Only one 3D structure is known (thermitase).

Proteinase K family

Large family of secreted endopeptidases found only in fungi, yeasts, and gram-negative bacteria as yet; the bacterial subgroup has >55% sequence identity. This family is characterized by a high degree of sequence similarity (>37% identity), only minor insertions and deletions and the absence of the Ca^{2+}-binding loop residues 76–81. Only a few of these enzymes have a significant C-terminal extension beyond the catalytic domain. One 3D structure is known (proteinase K).

Lantibiotic peptidase family

A small number of highly specialized enzymes for cleavage of leader peptides from precursors of lantibiotics, a unique group of post-translationally modified, antimicrobial peptides (Sahl et al., 1995). These endopeptidases have only been found in gram-positive bacteria, and several are intracellular. Only llnisp has a C-terminal extension, which acts as a membrane anchor. Characterized by low sequence similarity with each other and other subtilases (Fig. 3), and by numerous insertions/deletions. The most recently reported protein bspara from *Bacillus subtilis* is described as a putative protease required for plasmid stability; we speculate that it may also play a role in lantibiotic processing.

A few 3D structures have been predicted by homology modeling (Siezen et al., 1995a; Booth et al., 1996).

Kexin family

A large group of proprotein convertases (PCs) have been identified, all involved in activation of peptide hormones, growth factors, viral proteins, etc. (Barr, 1991; van de Ven et al., 1993). High specificity is seen for cleavage after dibasic (Lys-Arg or Arg-Arg) or multiple basic residues. Nearly all are eukaryotic and have high sequence homology (>40% identity), while two more distant members from *Aeromonas* and *Anabaena* provide links to other subti-

Fig. 2. (*facing page*) Alignment of amino acid sequences of catalytic domains of subtilases. Multiple sequence alignment was initially performed using the PILEUP program (Devereux et al., 1984). Next, improvements were made manually by taking into account the structure-based alignment (Siezen et al., 1991; Heringa et al., 1995). Inserts were judged to occur most likely in turns in external loops. Families A to F are indicated on the left. Enzyme acronyms are given in Table 1. (*) New entries, and (c) corrected entries since Siezen et al. (1991). Residue numbering at the top corresponds to that of mature subtilisin BPN′ (basbpn). Catalytic residues Asp 32, His 64, and Ser 221 are in bold (highlighted red), as is the oxyanion-hole residue Asn 155. Green = highly conserved residues from Table 4; yellow = Cys residues. Structurally conserved regions of the coreABC and extended coreAB are shown as solid bars; common secondary structure elements are shown as: h = helix, e = extended β-sheet, b = bend and t = β-turn (see also Fig. 1). The number of additional residues in large inserts in the catalytic domain, and in N- and C-terminal extensions, are shown in brackets. Each sequence begins at the mature N-terminus; an N-terminus based on the predicted pro-peptide cleavage site is indicated as (<). An unknown number of C-terminal residues is presented as (>). Residues 146–156 of bspara are from a different reading frame than proposed by the authors.

Table 1. *The subtilase family of serine proteases*

Acronym	Organism	cDNA,gene	Enzyme	Cellular Location	Amino acids			Signal Peptide	Accession code	
					Total	Prepro	Mature		EMBL	PDB
	BACTERIA									
	Gram-positive									
bss168	*Bacillus subtilis* 168	*aprA*	Subtilisin I168 (or E), aprA	Extra	381	106	275	+	K01988	+
basbpn	*Bacillus amyloliquefaciens*	*apr*	Subtilisin BPN;pr (NOVO)	Extra	382	107	275	+	X00165	2SNI
bssdy	*Bacillus subtilis* DY	–	Subtilisin DY	Extra	?	?	274	?	P00781	
blscar	*Bacillus licheniformis* NCIMB 6816	+	Subtilisin Carlsberg	Extra	379	105	274	+	X03341	1CSE
bls147	*Bacillus lentus*	+	Subtilisin 147, Esperase™	Extra	361	93	268	+	A08331	+
baalkp	*Bacillus alcalophilus* PB92	+	Subtilisin PB92	Extra	380	111	269	+	M65086	+
bsaprq	* *Bacillus* sp. NKS-21	*aprQ*	Subtilisin ALP I	Extra	374	102	272	+	D29736	
bsaprs	* *Bacillus* sp. G-825-6	*aprS*	Subtilisin Sendai	Extra	382	113	269	+	D29688	
bsta39	* *Bacillus* sp. TA39	+	Subtilisin TA39	Extra	420	111	309	+	X62369	
bsta41	* *Bacillus* sp. TA41	+	Subtilisin TA41	Extra	419	110	309	+	X63533	
bsspra	* *Bacillus* sp. LG12	*sprA*	Serine protease A	Extra?	>403	(>52)	(351)	+	U39230	
bssprb		*sprB*	Serine protease B	Extra?	824	(123)	(701)	+	U39230	
bssprc		*sprC*	Serine protease C	Extra	378	(103)	(275)	+	U39230	
bssprd		*sprD*	Serine protease D	Extra	379	(103)	(276)	+	U39230	
bseyab	*Bacillus subtilis* YaB	*ale*	Alkaline elastase YaB	Extra	378	110	268	+	M28537	
bssepr	* *Bacillus smithii*	+	Alkaline serine protease	Extra?	436	(100)	(336)	+	L24202	
bsepr	*Bacillus subtilis* 168	*epr*	Minor extracellular protease	Extra	645	103	542	+	X53307	
bsvpr	* *Bacillus subtilis* GP264	*vpr*	Minor extracellular protease	Extra	806	160	646	+	M76590	
bsbpf	*Bacillus subtilis*	*bpf*	Bacillopeptidase F	Extra	1,433	194	1,239	+	M29035	
bspara	* *Bacillus subtilis*	*parA*	Serine protease	Intra?	277	?	?	?	U55043	
bswpra	* *Bacillus subtilis 168*	*wprA*	Cell wall protease	Extra	894	(411)	(483)	+	U58981	
bsisp1	*Bacillus subtilis* IFO3013	*isp1*	Intracell. serine protease 1	Intra	319	(0)	(319)	–	M13760	
bsiakp	* *Bacillus* sp. 221	+	Intracell. alkaline protease	Intra	322	(0)	(322)	–	D10730	
bpisp	* *Bacillus polymyxa*	*isp*	Intracell. serine protease	Intra	326	(0)	(326)	–	D00862	
bsispq	* *Bacillus* sp. NKS-21	*ispQ*	Intracellular protease	Intra	323	(0)	(323)	–	D37921	
bsak1	* *Bacillus* sp. Ak.1	+	Proteinase Ak. 1	Extra	401	121	280	+	L29506	
tsiap	* *Thermoactinomyces sp.*	*tiap*	Intracell. alkaline protease	Intra	321	(0)	(321)	–	D87557	
tstap	* *Thermoactinomyces* sp. E79	+	Thermostable alkaline protease	Extra	384	106	268	+	U31759	
tvther	*Thermoactinomyces vulgaris*	–	Thermitase	Extra	?	?	279	?	P04072	2TEC
efcyla	*Enterococcus faecalis*	*cylA*	Cytolysin component A	Extra	412	95	317	+	M38052	
spscpa	*Streptococcus pyogenes*	*scpA*	C5a peptidase	Extra	1,167	(31)	(1,136)	+	J05229	
seepip	*Staphylococcus epidermidis* Tu3298	*epiP*	Epidermin leader peptidase	Extra	461	?	?	+	X62386	
sepepp	* *Staphylococcus epidermidis* 5	*pepP*	Pep5 leader peptidase	Intra?	285	(0)	(285)	–	Z49865	
seelkp	* *Staphylococcus epidermidis* K7	*elkP*	Epilancin leader peptidase	Intra?	>130	(0)	?	–	U20348	
llnisp	* *Lactococcus lactis* NIZO R5	*nisP*	Nisin leader peptidase	Extra	682	?	?	+	L11061	
llprtp	*Lactococcus lactis* (cremoris) SK11	*prtP*	Cell-envelope proteinase	Extra	1,962	187	1,775	+	J04962	
ldprtb	* *Lactobacillus delbrueckii* (*bulgaricus*)	*prtB*	Cell-envelope proteinase	Extra	1,946	(192)	(1,754)	+	L48487	
lslasp	* *Lactobacillus sake*	*lasP*	Lactocin S leader peptidase	Intra?	266	(0)	(266)	–	Z54312	

	Gram-negative									
dnbpr	*Dichelobacter nodosus*	*bpr*	Basic protease	Extra	603	132	344	+	Z16080	
dnapv5	* *Dichelobacter nodosus*	*aprV5*	Acid protease V5	Extra	595	130	347	+	L18984	
dnapv2	* *Dichelobacter nodosus*	*aprV2*	Acid protease V2	Extra	599	127	345	+	L38395	
xcproa	*Xanthomonas campestris*	+	Extracellular protease	Extra	580	(136)	(444)	+	X51635	
smserp	*Serratia marcescens* IFO3046	+	Extracellular serine protease	Extra	1,045	27	(381)	+	M13469	
smssp1	* *Serratia marcescens*	+	SSP-h1 protease	Extra	1,036	(42)	(994)	+	D78380	
smssp2	* *Serratia marcescens*	+	SSP-h2 protease	Extra	1,034	(35)	(999)	+	D78380	
taaqua	*Thermus aquaticus* YT-1	*pstI*	Aqualysin I	Extra	513	127	281	+	X07734	
trt41a	c *Thermus* rT41A	*aprA*	T41A protease	Extra	408	130	278	+	U17342	
alapr1	* *Alteromonas* sp. 0-7	*aprI*	Serine protease	?	729	(150)	(579)	+	D38600	
alapr2	* *Alteromonas* sp. 0-7	*aprII*	Alkaline serine protease	Extra	621	120	401	+	S68495	
vaproa	*Vibrio alginolyticus*	*proA*	Protease A	Extra	534	(141)	(393)	+	M25499	
vmvapt	* *Vibrio metschnikovii*	*vapT*	Alkaline serine protease	Extra	547	119	428	+	Z28354	
asaspa	* *Aeromonas salmonicida*	*aspA*	Serine protease	Extra	621	24	597	+	X67043	
phssa1	* *Pasteurella haemolytica*	*ssa1*	Serotype-specific antigen	Extra	932	(26)	(906)	+	P31631	
psaprp	* *Pseudomonas* sp. KFCC10818	*aprP*	Extracellular serine protease	Extra	422	139	283	+	U36429	
slssp	* *Streptomyces lividans* 66	*ssp*	Protease/aminopeptidase	Extra	513	124	389	+	L41655	
scsepr	* *Streptomyces coelicolor*	+	Serine protease	?	390	?	?	?	U33176	
	Cyanobacteria									
avprca	c *Anabaena variabilis*	*prcA*	Ca-dependent protease	Intra	620	(191)	(419)	?	X56955	
sy0535	* *Synechocystis* sp.	*0535*	Putative serine protease	?	613	(111)	(502)	+	D64006	
	ARCHAEA									
nahlys	* *Natriaba asiatica* 172 P1	*hly*	Halolysin 172 P1	Extra	530	119	411	+	D01201	
hmhlys	* *Haloferax mediterranei* R4	*hlyR4*	Halolysin R4	Extra	519	116	403	+	D64073	
paalys	* *Pyrobaculum aerophilum*	+	Aerolysin	?	>397	(>90)	(307)	?	S76079	
pfpyro	* *Pyrococcus furiosus*	+	Pyrolysin	Extra	1,398	149	1249	+	U55835	
tsplst	* *Thermococcus stetteri*	+	Heat-stable protease	Extra	>830	(159)	>680	+	[a]	
smstab	* *Staphylothermus marinus*	+	STABLE protease	Extra	1,345	(182)	(1,163)	+	U57968	
	EUKARYA									
	Fungi									
taprok	*Tritirachium album* Limber	+	Proteinase K	Extra	384	105	279	+	X14689	2PRK
tapror	*Tritirachium album*	+	Proteinase R	Extra	387	108	279	+	X56116	
taprot	*Tritirachium album*	*proT*	Proteinase T	Extra	>293	?	281	?	M54901	
bbpr1	* *Beauveria bassiana*	+	Protease Pr1	Extra	(360)	99	(261)	+	U16305	
fusalp	* *Fusarium* sp. S-19-5	*alp*	Alkaline protease	Extra	379	99	280	+	S71812	
macdpa	* *Metarhizium anisopliae*	*pr1*	Cuticle-degrading protease	Extra	388	107	281	+	M73795	
anpepc	* *Aspergillus niger*	*pepC*	Serine protease	Extra	533	(136)	(497)	+	M96758	
plbspr	* *Paecilomyces lilacinus*	+	Serine protease	?	>367	(>83)	(284)	?	L29262	
anpepd	* *Aspergillus niger*	*pepD*	Alkaline protease	Extra	416	(121)	(295)	+	L19059	
anprta	* *Aspergillus nidulans*	*prtA*	Alkaline protease	Extra	403	(120)	(283)	+	L31778	
aooryz	*Aspergillus oryzae* ATCC20386	+	Alkaline protease	Extra	403	121	282	+	D00350	
aforyz	* *Aspergillus fumigatus*	*alp*	Alkaline protease	Extra	403	121	282	+	Z11580	

(continued)

Table 1. *Continued*

Acronym	Organism	cDNA,gene	Enzyme	Cellular Location	Amino acids			Signal Peptide	Accession code	
					Total	Prepro	Mature		EMBL	PDB
afelst	* *Aspergillus flavus* 28	+	Elastinolytic proteinase	Extra	403	121	282	+	L08473	
acalpr	*Acremonium chrysogenum*	*alp*	Alkaline protease	Extra	402	120	282	+	D00923	
thprb1	* *Trichoderma harzianum*	*prb1*	Alkaline protease	Extra	409	120	289	+	M87518	
aoespr	* *Arthrobotrys oligospora*	*PII*	PII protease	Extra	408	(122)	(286)	+	X94121	
	Slime molds									
ddtagb	* *Dictyostelium discoideum*	*tagB*	Serine protease/transporter	?	1,905	?	?	+	U20432	
ddtagc	* *Dictyostelium discoideum*	*tagC*	Serine protease/transporter	?	1,744	?	?	+	U60086	
	Yeasts									
klkex1	*Kluyveromyces lactis*	*kex1*	Kex1 serine proteinase	Golgi	700	(102)	(598)	?	X07038	
spsepr	* *Schizosaccharomyces pombe*	+	Serine protease	?	467	(177)	(290)	+	D14063	
spkrp1	* *Schizosaccharomyces pombe*	*krp*	Dibasic endopeptidase	?	709	(102)	(607)	+	X82435	
sckex2	*Saccharomyces cerevisiae*	*kex2*	Kex2 serine proteinase	Golgi	814	109	705	+	M24201	
scprb1	*Saccharomyces cerevisiae*	*prb1*	Protease B, cerevisin	Vacuole	635	280	(355)	+	M18097	
scysp3	* *Saccharomyces cerevisiae*	*ysp3*	Protease III	?	478	(168)	(310)	+	Z74911	
scyct5	* *Saccharomyces cerevisiae*	*yct5*	Proteinase (hypothetical)	?	491	(91)	(400)	+	X59720	
ylxpr2	*Yarrowia lipolytica*	*xpr2*	Alkaline extracellular protease	Extra	454	157	297	+	M23353	
ylxpr6	* *Yarrowia lipolytica*	*xpr6*	Dibasic processing protease	Golgi	976	(171)	(805)	+	L16238	
	Plants									
llsp09	* *Lilium longiflorum*	*lim9*	Serine proteinase	?	(795)	(114)	(681)	+	D21815	
atserp	* *Arabidopsis thaliana*	*ara12*	Serine proteinase	?	>746	(>97)	(649)	?	X85974	
agserp	* *Alnus glutinosa*	*ag12*	Serine proteinase	?	761	(113)	(648)	+	X85975	
lep69	* *Lycopersicon esculentum*	+	P69 protease	?	745	114	631	+	X95270	
cmcucu	* *Cucumis melo*	+	Cucumisin	?	731	110	621	+	D32206	
paaf70	* *Picea abies*	+	Antifreeze protein af70	?	779	(107)	(672)	+	D86598	
	Insects									
dmfur1	*Drosophila melanogaster*	*fur1*	Furin 1	?	892	(309)	(583)	?	X59384	
dmfur2	* *Drosophila melanogaster*	*fur2*	Furin 2	?	1,680	(319)	(1,361)	?	M94375	
dmpga9	* *Drosophila melanogaster*	*pga9*	Tripeptidase (hypothetical)	?	>826	?	?	?	[b]	
sffur	* *Spodoptera frugiperda*	+	Furin	?	1,299	(104)	(1,195)	+	Z68888	
aafur	* *Aedes aegypti*	+	Furin	?	1,060	(215)	(845)	+	L46373	
	Coelanterata									
hakx2a	* Hydra attenuata	+	Kex2-like endoprotease	?	793	(152)	(641)	+	M95931	
	Nematoda									
cefur2	* *Caenorhabditis elegans*	*bli4*	Blisterase A	?	684	(116)	(568)	+	L29438	
cepc2	* *Caenorhabditis elegans*	*pc2*	PC2 protease	?	652	(107)	(545)	+	U04995	
cefur1	* *Caenorhabditis elegans*	+	Kex2-like endoprotease	?	681	(128)	(553)	+	U12682	
cetpp	* *Caenorhabditis elegans*	+	Tripeptidase	Intra?	1,374	?	?	−	U23176	

	Mollusca								
lspc2	* *Lymnea stagnalis*	+	PC2 protease	?	653	(116)	(537)	+	X68850
lsfur2	* *Lymnea stagnalis*	*lfur2*	PCX protease	?	837	(113)	(724)	+	S69833
acfur1	* *Aplysia californica*	*fur*	Furin	?	705	(104)	(601)	+	L28769
acpc2	* *Aplysia californica*	+	PC2 protease	?	653	(116)	(537)	+	X59682
acfur2	* *Aplysia californica*	*pc1A*	PCX protease (furin 2)	?	824	(95)	(729)	+	L38715
acpc1	* *Aplysia californica*	*pc1B*	PC1 protease	?	712	(119)	(593)	+	L28767
	Arthropoda								
ttfur	* *Tachypleus tridentatus*	+	Kex2-like endoprotease	?	752	(112)	(640)	+	D83994
	Amphibia								
xlfura	* *Xenopus laevis*	*xen14*	Furin A	?	783	(105)	(678)	+	M80471
xlpc2	* *Xenopus laevis*	*xy2.1*	PC2 protease	?	639	(110)	(529)	+	X66493
	Fish								
bcpc2	* *Branchiostoma californiensis*	+	PC2 protease	?	689	(114)	(575)	+	U22051
bpcp3	* *Branchiostoma californiensis*	+	PC3 protease	?	774	(110)	(664)	+	U22052
lapc1	* *Lophius americanus*	+	PC1 protease	?	775	(113)	(662)	+	U01910
	Mammals								
hsfur	Human	*fur*	Furin	?	794	107	687	+	X17094
hspac4	Human	+	PACE4 protease	?	969	(149)	(820)	+	M80482
hspc13	* Human	*nec1*	PC1/PC3 protease	?	753	110	643	+	X64810
hspc2	* Human	*nec2*	PC2 protease	?	638	109	529	+	J05252
mmpc4	* Mouse	+	PC4 protease	?	655	(110)	(545)	+	D01093
hspc56	* Human	+	PC5/6 protease	?	915	(116)	(799)	+	U56387
hslpc	* Human	+	Lymphoma protease, PC7	?	785	(141)	(644)	+	U33849
hstpp2	c Human	+	Tripeptidyl peptidase II	Intra	1,249	(0)	(1,249)	−	M73047
hskiaa	* Human	+	Putative serine protease	?	1,052	?	?	+	D42053
	VIRUSES								
hvccvp	* Herpesvirus 1	*orf47*	Catfish virus protease	?	518	?	?	−	M75136

*new; c, corrected.
[a]Voorhorst et al., in prep.
[b]R. Nusse, F. Van Leeuwen, pers. comm.

Table 2. *Variants of subtilases with >90% sequence identity in catalytic domain*

Main	Variant	Organism	cDNA,gene	Enzyme	Amino acid differences	Accession codes	
						EMBL	PDB
bss168		*Bacillus subtilis* 168	aprA	Subtilisin I168 (or E), aprA		K01988	+
	bssas	*Bacillus subtilis* (*amylosacchariticus*)	+	Subtilisin Amylosacchariticus	2	D00264	
	bsaprj	*Bacillus stearothermophilus* NCIMB 10278	aprJ	Subtilisin J	2	M64743	
	bsaprn	*Bacillus subtilis natto* NC2-1	aprN	Subtilisin NAT	2	D25319	
	bmsamp	*Bacillus mesentericus*	+	Mesentericopeptidase	5	P07518	1MEE
blscar		*Bacillus licheniformis* NCIMB 6816	+	Subtilisin Carlsberg		X03341	1CSE
	blkera	*Bacillus licheniformis* PWD-1	kerA	Keratinase	4	S78160	
	blsca3	*Bacillus licheniformis* 15413	subC	Subtilisin Carlsberg 15413	8	X91262	
	blsca2	*Bacillus licheniformis* 14353	subC	Subtilisin Carlsberg 14353	15	X91261	
	blsca1	*Bacillus licheniformis* 11594	subC	Subtilisin Carlsberg 11594	16	X91260	
baalkp		*Bacillus alcalophilus* PB92	+	Subtilisin PB92, Maxacal™		M65086	+
	blsavi	*Bacillus lentus*	+	Savinase™	1	P29600	
	bsksmk	*Bacillus* sp. KSM-K16	+	M-protease	4	Q99405	1MPT
	blsubl	*Bacillus lentus*	−	Subtilisin BL	6	P29599	1ST3
bls147		*Bacillus lentus*	+	Subtilisin 147, Esperase™		A08331	+
	bsaprm	*Bacillus* sp. B18′	aprM	Subtilisin AprM	3	D26542	
	bah101	*Bacillus* sp. AH-101	+	Alkaline protease AH-101	9	D13158	
bsbpf		*Bacillus subtilis*	bpf	Bacillopeptidase F		M29035	
	bshspk	*Bacillus subtilis* (natto)16	hspK	Bacillopeptidase F	3	D44498	
spscpa		*Streptococcus pyogenes*	scpA	C5a peptidase		J05229	
	spm49	*Streptococcus pyogenes* M49	scpA49	C5a peptidase	19	X78055	
	sscpb	*Streptococcus* sp. 78-471	scpB	C5a peptidase	13	U56908	
llprtp		*Lactococcus lactis* (*cremoris*) SK11	prtP	PIII-type proteinase		J04962	
	llwg2	*Lactococcus lactis* Wg2	prtP	PI-type proteinase	18	M24767	
	ll763	*Lactococcus lactis* NCDO763	prt	PII-type proteinase	22	X14130	
	lpprtp	*Lactobacillus paracasei* NCDO151	prtP	PII-type proteinase	31	M83946	
dnbpr		*Dichelobacter nodosus* 198	bpr	Basic protease		Z16080	
	dnbprb	*Dichelobacter nodosus* 305	bprB	Basic protease	21	L37754	
avprca		*Anabaena variabilis*	prcA	Ca-dependent protease		X56955	
	asprca	*Anabaena* sp. PCC7120	prcA	Ca-dependent protease	8	X63439	
hstpp2		Human	+	Tripeptidyl peptidase II		M73047	
	mmtpp2	Mouse	+	Tripeptidyl peptidase II	17	X81323	
	rntpp2	Rat	+	Tripeptidyl peptidase II	15	U50194	
hsfur		Human	fur	Furin		X17094	
	cgfur	Hamster	fur	Furin	3	U20436	
	mmfur	Mouse	fur	Furin	3	X54056	

	rnfur	Rat	+	Furin	3	X55660
	btfur	Cow	+	Furin	6	X75956
	ggfur	Chicken	+	Furin	20	Z68093
hspc13		Human	nec1	PC1/PC3, neuroendocrine convertase 1		X64810
	mmpc13	Mouse	nec1	PC1/PC3	8	M69196
	rnpc13	Rat	+	PC1/PC3	4	M76705
	sspc13	Pig	+	PC1/PC3	3	U20545
hspc2		Human	nec2	PC2, neuroendocrine convertase 2		J05252
	mmpc2	Mouse	nec2	PC2, neuroendocrine convertase 2	7	M55669
	rnpc2	Rat	rpc2	PC2, neuroendocrine convertase 2	9	M76076
	sspc2	Pig		PC2, neuroendocrine convertase 2	3	X68603
lspc2		*Lymnea stagnalis*	+	PC2		X68850
	acpc2	*Aplysia californica*	+	PC2	20	X59682
hspac4		Human	+	PACE4		M80482
	mmpac4	Mouse	+	PACE4	10	D50060
	rnpac4	Rat	+	PC7	10	L31894
mmpc4		Mouse	+	PC4, proprotein convertase		D01093
	rrpc4	Rat	pc4	PC4, proprotein convertase	3	L14937
hspc56		Human	+	PC5/6		U56387
	mmpc56	Mouse	+	PC5/6, proprotein convertase	2	L14932
	rrpc56	Rat	+	PC5, proprotein convertase	3	L14933
xlfura		*Xenopus laevis*	xen14	Furin A		M80471
	xlfurb	*Xenopus laevis*	xen18	Furin B	7	M80472
dmfur2		*Drosophila melanogaster*	fur2	Furin 2		M94375
	sffur	*Spodoptera frugiperda*	+	Furin	27	Z68888
hslpc		Human	+	Lymphoma protease, PC8		U33849
	rnpc7	Rat	+	PC7	8	U36580
	mmpc7	Mouse	+	PC7	12	U48830
	xllpc	*Xenopus laevis*	+	lpc	31	[a]

[a]G. Matthews, pers. comm.

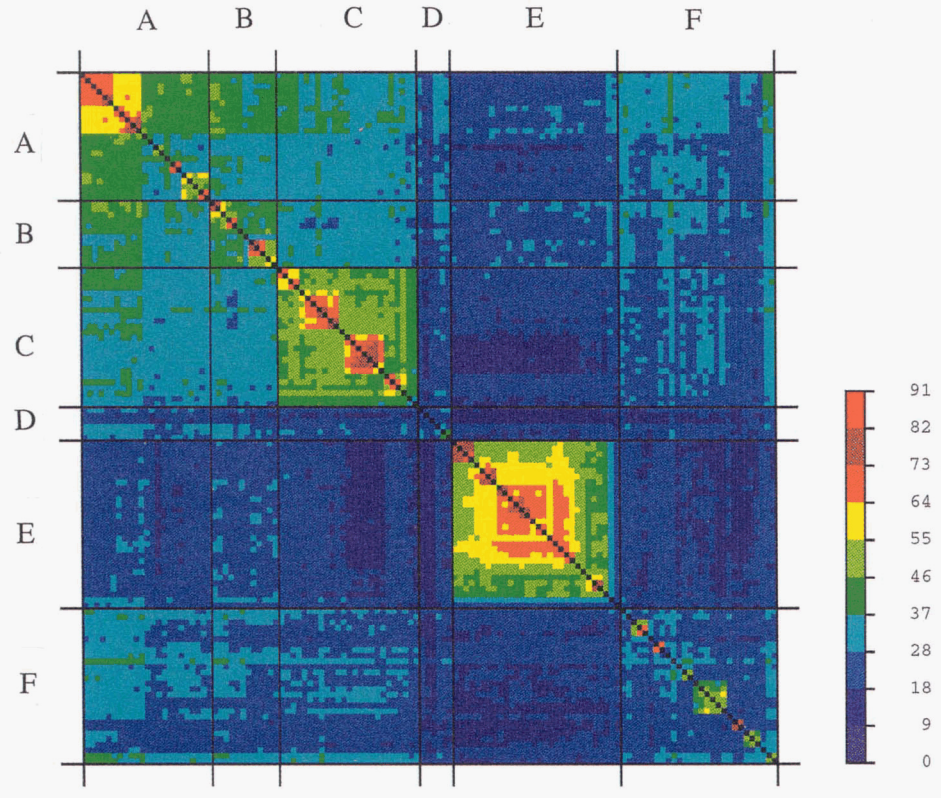

Fig. 3. Pair-wise sequence identity matrix. Sequences are plotted vertically and horizontally in the same order as in Figure 2; the incomplete sequence of hvccvp is not included. Subdivision into families A to F is indicated. A color code bar for percentage sequence identity is shown.

lase families. A subgroup of yeast enzymes is evident, as are subgroups of PC1 (>55% identity), PC2 (>73% identity), and furin (>55% identity). In catfish herpes virus 1 a related but incomplete amino acid sequence has been found that is presumed to have been captured from a host (Rawlings and Barrett, 1994).

Several 3D structures have been predicted by modeling (see below).

Pyrolysin family

Heterogeneous group of enzymes of varied origin and low sequence conservation (most <37% identity). Characterized by large insertions and/or long C-terminal extensions, many with sequence homology suggesting common ancestors. The most extreme example is llsp09 from the plant *Lilium* with insertions totaling more than 260 residues compared to subtilisin, almost doubling the size of the catalytic domain. Subgroups of tripeptidyl peptidases and plant subtilases (>37% identity) are distinguished; the former are of higher eukaryotic origin, but only the human and mouse enzymes have actually been identified biochemically as tripeptidyl peptidases.

Several 3D structures have been predicted by modeling (see below).

Several other subtilases have been identified for which only the N-terminal or other partial sequence of the purified enzyme is available; based on sequence alignment with Figure 2, these subtilases presumably belong to families A, B, and C (Table 3).

Conserved residues

Highly conserved residues are listed in Table 4 and highlighted in Figure 2. Only the essential catalytic triad residues D32, H64, and S221 and a single glycine residue (G219) are totally conserved in all sequences. Four other glycine residues (34, 65, 83, and 154) are varied only once or twice; G34 and G154 have main-chain torsion angles that do not allow for amino acid residues with side chains. At several other positions the variation is limited to two or three residues, which are usually structurally similar. In general, the residues of the two internal helices hC and hF are the most highly conserved in all subtilases. Three amino acid sequences (lslasp, sepepp, and asaspa) are particularly poorly conserved; although it seems questionable whether these enzymes are functional, a mutation analysis of the *pepP* gene suggests that it indeed encodes a functional protease (Meyer et al., 1995).

Many more residues are totally conserved within each of the six families A to F, and these can be used to identify new family members. In particular, families A and C are most conserved, with a total of 32 and 41 invariant residues, respectively, while family E has 63 invariant residues if the two more divergent sequences (asaspa and avprca) are excluded.

Residue N155 (in a conserved segment 152–155), which helps to stabilize the oxyanion generated in the tetrahedral transition state (Carter and Wells, 1990), is not fully conserved. The only accepted substitution here is N155D, as is found in the PC2 subgroup of the kexin family. The effect of this substitution on the

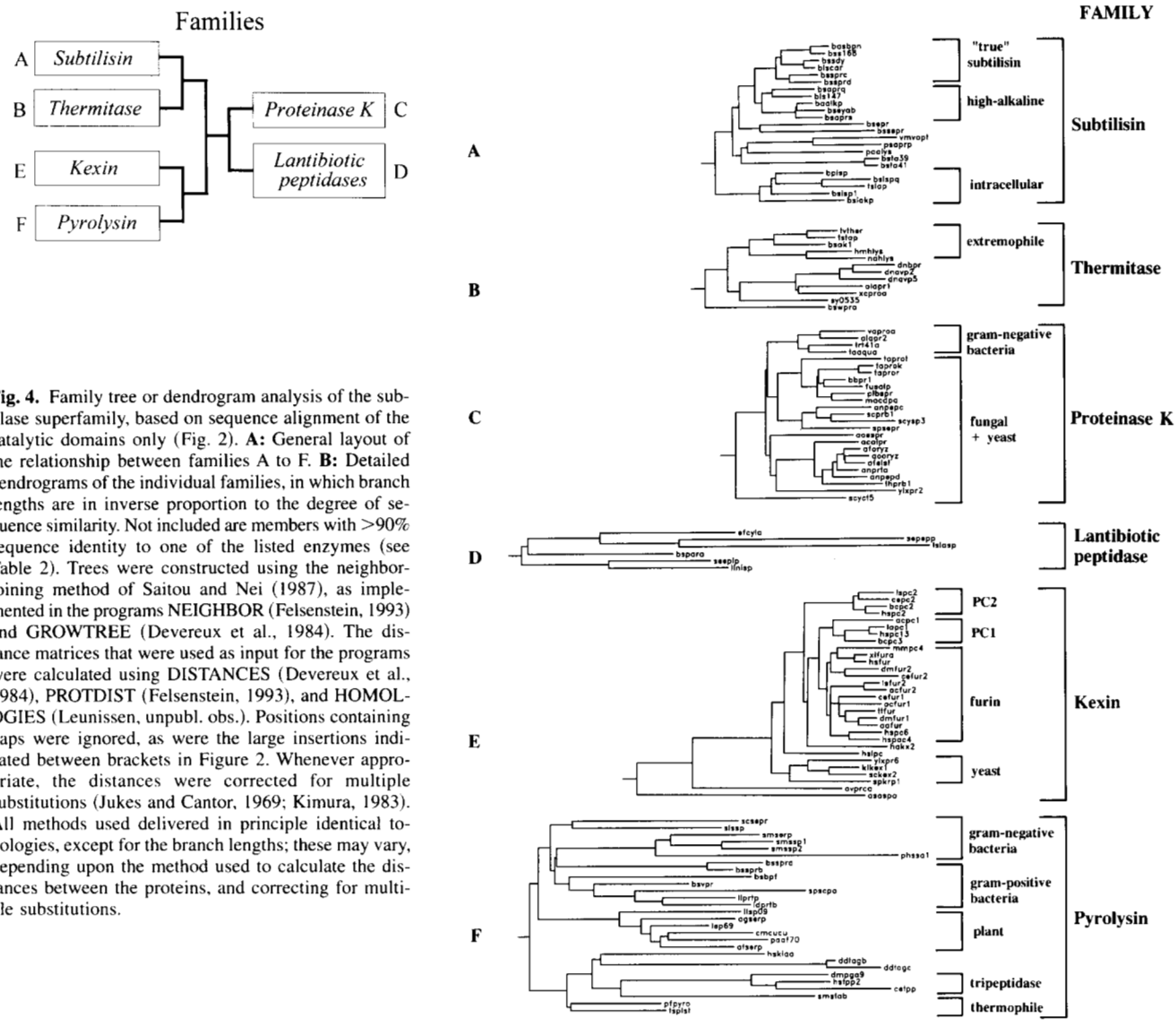

Fig. 4. Family tree or dendrogram analysis of the subtilase superfamily, based on sequence alignment of the catalytic domains only (Fig. 2). **A:** General layout of the relationship between families A to F. **B:** Detailed dendrograms of the individual families, in which branch lengths are in inverse proportion to the degree of sequence similarity. Not included are members with >90% sequence identity to one of the listed enzymes (see Table 2). Trees were constructed using the neighbor-joining method of Saitou and Nei (1987), as implemented in the programs NEIGHBOR (Felsenstein, 1993) and GROWTREE (Devereux et al., 1984). The distance matrices that were used as input for the programs were calculated using DISTANCES (Devereux et al., 1984), PROTDIST (Felsenstein, 1993), and HOMOLOGIES (Leunissen, unpubl. obs.). Positions containing gaps were ignored, as were the large insertions indicated between brackets in Figure 2. Whenever appropriate, the distances were corrected for multiple substitutions (Jukes and Cantor, 1969; Kimura, 1983). All methods used delivered in principle identical topologies, except for the branch lengths; these may vary, depending upon the method used to calculate the distances between the proteins, and correcting for multiple substitutions.

catalytic efficiency of these proteases has been investigated by protein engineering (Benjannet et al., 1995; Zhou et al., 1995).

Homology modeling

The procedure for homology modeling and protein engineering of the catalytic domain of subtilases of unknown 3D structure based on known crystal structures was described in our previous review (Siezen et al., 1991), and can be applied to any of the enzymes listed in Tables 1 and 2.

Modeling should be based on the known crystal structure of the most related enzyme, and this will be straightforward for members of the families A–C, because 3D structures are known in each family. For the families D–F, with no known 3D structures, modeling will be less straightforward and can be based on any known structure from families A–C or a combination of these. Problems will arise where large insertions occur, because these are still impossible to model reliably. It would be extremely helpful for modeling purposes to determine the crystal structure of at least one member of each of the D–F families, preferably those with large inserts.

This homology method has since been refined and applied for modeling and engineering of (a) the cell-envelope proteinase llprtp of *Lactococcus lactis* (Siezen et al., 1993; Bruinenberg et al., 1994a, 1994b); (b) the lantibiotic leader peptidases llnisp of *Lactococcus lactis* (Van der Meer et al., 1994; Siezen et al., 1995a), and efcyla of *Enterococcus faecalis* (Booth et al., 1996); (c) the kexin family members furin (hsfur: Creemers et al., 1993; Siezen et al., 1994) and PC2/PC3 (Lipkind et al., 1995); and (d) the heat-stable proteases pfpyro and tsplst of the hyperthermophiles *Pyrococcus furiosus* and *Thermococcus stetteri* (W. Voorhorst, A. Warner, W. de Vos, R. Siezen, in prep.). These studies have provided predictions and evidence for inserted and disposable loops, disulfide bridges, Ca^{2+}-ion binding sites, surface salt bridges and networks, aromatic surface clusters, and residues involved in enzyme–substrate interactions. Some examples are discussed below.

Table 3. *Incomplete amino acid sequences of subtilases*

Organism	Enzyme	Acronym	Residues determined		Family	References
			N-term.	Other		
BACTERIA						
Gram-positive						
Bacillus subtilis A50	Intracell. serine protease	bsia50	1–54		A	Strongin et al., 1978
Bacillus sp. GX6644	Subtilisin GX	bssugx	1–16		A	Durham, 1993
Bacillus sp. Y	Protease BYA	bspbya	1–21		A	Shimogaki et al., 1991
Bacillus thuringiensis israelensis	Extracellular serine protease	btisra	1–14	223–243	B	Chestukhina et al., 1986
Bacillus thuringiensis finitimus	Extracellular serine protease	btfini	1–15		B	Chestukhina et al., 1986
Bacillus thuringiensis kurstaki	Extracellular serine protease	btkurs	6–20		B	Kunitate et al., 1989
Bacillus cereus	Extracellular serine protease	bcespr	1–15	223–243	B	Chestukhina et al., 1986
Bacillus intermedius 3–19	Alkaline serine protease	biprot	1–15		A	Balaban et al., 1994
Nocardiopsis dassonvillei (prasina)	Alkaline serine protease	ndapII	1–26		C	Tsujibo et al., 1990
Gram-negative						
Streptomyces rutgersensis	Proteinase D	srespd	1–23		C	Lavrenova et al., 1984
Thermus Tok3A1	Caldolysin	tscald	1–15		C	Freeman et al., 1993
Vibrio metschnikovii	Alkaline protease VapK	vmapk	1–36		A	Kwon et al., 1994
Cochliobolus carbonum	Extracellular protease	ccalp2	1–29		C	Murphy & Walton, 1996
EUKARYA						
Fungi						
Agaricus bisporus	Extracellular serine protease	abexpr	1–19		C	Burton et al., 1993
Malbranchea sulfurea	Thermomycolin	msthmy	1–28	217–222	C	Gaucher & Stevenson, 1976
Ophiostoma piceae	Extracellular protease	opexpr	1–18	170–193	C	Abraham & Breuil, 1995
Verticillium chlamydosporium	Extracellular protease VCP1	vcexp1	1–20		C	Segers et al., 1995
Scedosporium apiospermum	Extracellular protease	saalpr	1–13		C	Larcher et al., 1996

Table 4. *Highest conserved residues in subtilases (v = variability)*

Residue	v = 1	v = 2	v = 3	Context/function	Exception
32	D			Catalytic triad residue	
34	G			Bend; ϕ, ψ = 99°, 179°	N (lslasp), A (smserp), P (smsspl, smssp2)
64	H			Catalytic triad residue	
65	G			Buried helix, close packing	del (asaspa)
68			V,C,T	Buried helix, close packing, directly under catalytic triad	M (lslasp), I (sepepp, ddtagc)
69			A,S,C	Buried helix, close packing	G (nahlys), T (bsbpf), I (sepepp)
70		G,S		Buried helix, close packing	T (smstab), A (smssp1, paaf70)
83	G			Helix/turn, close packing	A (lslasp), T (efcyla)
90			L,I,V	Buried β-strand, hydrophobic packing to helix C	W (bsbpf), M (seepip)
125	S			Bend, directly adjacent to catalytic triad	P (lslasp), C (hakx2), T (acfur1, bcpc2)
152		A,S		Lines S1 pocket	
154	G			Lines S1 pocket; ϕ, ψ = 114°, 163°	M (sepepp), del (bssepr)
155	N			Oxyanion stabilization	D (lspc2, bcpc2, cepc2, hspc2)
189			F,Y,W	Turn at surface, side chain turned into pocket	del (sepepp), S (smserp), L (bspara)
193			G,C,N	Begin turn	Y (sepepp), D (dmpga9), T (vmvapt)
201			P,Y,F	Bend at end β-strand, hydrophobic ring stacks with H226	I (seepip, smstab)
219	G			Bend between e9 and hF; ϕ, ψ = 147°, 160°	
220	T			OD1 H-bonded to backbone NH-154	N (sepepp, llnisp)
221	S			Catalytic triad residue	
223		A,S		Buried helix, close packing	
225	P			Buried helix, close packing	G (lslasp), S (sepepp, ddtagc)
229		G,A		Buried helix, close packing	T (bssepr)

Large insertions and deletions

The 190 residues that constitute the scrs, as defined from the known crystal structures (Siezen et al., 1991) and shown in Figure 2 are present in nearly all the subtilases. Some unusual deletions are found, however, as listed in Table 5, and this implies that not all of these core residues are essential for proper folding. Most of these deletions occur in subtilase family D, the lantibiotic leader peptidases, and include large N- and C-terminal deletions. All but one of the internal deletions can be readily accommodated by connecting residues that are spatially adjacent in the 3D structures of subtilisin/thermitase. Particularly interesting in this respect is the natural deletion of the Ca1-ion binding loop, residues 74–82, in the *Enterococcus* subtilase (efcyla), thereby presumably extending helix C by another four residues (Booth et al., 1996); this is precisely the loop deletion that was engineered into subtilisin to

Table 5. *Large or unusual deletions and insertions*

Unusual deletion			
Missing residues	Context	Family	Enzyme
1–13	N-terminus, hA	D	sepepp
65–66	Part hC, adjacent catalytic His	E	asaspa
74–82	Ca-binding loop → hC extended	D	efcyla
96–102	Turn, substrate-binding region	F	smserp
180–189	Turns	D	sepepp
257–275	C-terminus, hH	D	lslasp

Large insertion				
Inserted residues				
Position	Number	Properties	Family	Enzyme
N-term.	Up to 98	No homology	E	Most family members
	59	Highly charged	F	spscpa
	34	Highly charged	C	scyct5
vr1	18		C	scyct5
vr4	30–33	Weak homology	F	spscpa, llprtp, ldprtb
	28–30	High homology	B	dnbpr, dnavp2, dnavp5, xcproa, alaprl
	26–31	Medium homology, conserved S–S bond ?	F	llsp09, atserp, cmcucu, agserp, lep69, paaf70
	23	High homology	F	smsspl, smssp2
vr5	147–213	Weak homology, see alignment in Fig. 5	F	pfpyro, tsplst, dmpga9, hstpp2, cetpp
vr6	30		F	pfpyro
vr7	42	Highly charged (50%)	F	phssa1
vr8	16	Highly charged	C	anpepC
vr9	51		F	smserp
	34	High homology	F	smsspl, smssp2
	18		F	phssa1
vr11	22		F	slssp
	16–18	Weak homology	D	seepip, llnisp
vr13	134–169	Weak homology in central section (Fig. 5)	F	spscpa, llprtp, ldprtb, bsvpr, lep69, paaf70, atserp, agserp, cmcucu, llsp09
	73–75	High homology	F	ddtagb,ddtagc
	27		F	pfpyro
vr15	20–22	Weak homology	D	efcyla, seepip, llnisp
vr16	149		A	vmvapt
	27	S–S bond ?	E	asaspa
	22		F	smserp
	20	High homology	F	bsspra, bssprb
vr18	19		B	sy0535
vr19	38	S–S bond ?	E	cepc2
	34		E	asaspa
	25		B	bswpra
	22–24	High homology	F	ddtagb, ddtagc
	21	S–S bond?	F	slssp

obtain a Ca^{2+}-independent, faster-folding variant (Gallagher et al., 1993, 1995). The only unusual deletion comprises residues 65–66 of helix C adjacent to the catalytic His; this deletion is not due to a sequencing error (G. Coleman, pers. comm.).

The vrs essentially comprise all the connections between conserved elements of secondary structure, as shown schematically in Figure 1. While the positions of vrs are essentially the same as defined before (Siezen et al., 1991), the length of these vrs is now found to vary considerably more. The largest and most unusual insertions are listed in Table 5. Some of these large inserts are unique for a single enzyme, for example, pfpyro (in vr6), phssa1 (in vr7), vmvapt (in vr16), and cepc2 (in vr19). Large insertions occur most frequently in vr5, vr13, and vr16. Sequence conservation in large inserts is frequently also apparent, particularly within subtilase family B (inserts in vr4), as shown in Figure 2, and within family F, as shown in the alignments in Figure 5. This is further evidence for a common evolutionary origin of subgroups of enzymes that were already clustered together in the cladogram (Fig. 4). Also note that the inserts in the kexin family E are not large, but they are highly conserved (Fig. 2).

```
vr5
                                                                    total
dmpga9  -(10)-GNIKGLSGNSLKLS-(103)-YDCILFPTADGWLTIVDTTEQGDL-(38)-   188
hstpp2  -( 9)-GEIVGLSGRVLKIP-( 95)-YDCLVWHDGEVWRACIDSNEDGDL-(40)-   181
cetpp   -( 9)-GVIEGISGRKLAIP-( 96)-ADVVTWHDGEMWRVCIDTSFRGRL-(40)-   182

vr13
                                                                    total
11sp09  -(62)-VRGKLIICTLTTDSSSPMSIEAILSTIQKIGAVGVIITMD-(67)-        169
agserp  -(51)-VY-SVVICEAITP------IYAQIDAITRSNVAGAILISN-(51)-        135
cmcucu  -(59)-LKGKIVVCEASFG---------HEFFKSLDGAAGVLMTSN-(50)-        140
paaf70  -(62)-AKGNVVVCIANDTAA---SRYIMKLAVQDAGGIGMVVVED-(50)-        149
lep69   -(59)-IRGKIVICLAGGG----VPRVDKGQAVKDAGGVGMIIINQ-(52)-        147
atserp  -(58)-VKGKIVMCDRGIN-----ARVQKGDVVKAAGGVGMILANT-(52)-        145
bsvpr   -(51)-LTGKVAVVKRGSI-----AFVDKADNAKKAGAIGMVVYNN-(48)-        134
spscpa  -(50)-VKGKIALIERGDI-----DFKDKVANAKKAGAVGVLIYDN-(49)-        134
11prtp  -(60)-AKGKIAIVKRGEF-----SFDDKQKYAQAAGAAGLIIVNT-(56)-        151
ldprtb  -(64)-VKGQLAVVKRGAY-----TFSAKVANAKAAGAAGIVIYNS-(51)-        150

consensus      *kgk****  g          k   *  aga*G****
```

Fig. 5. Sequence alignment of most homologous regions in large inserts in vr5 and vr13. The numbers of additional residues in these large inserts are shown in brackets. Consensus residues are indicated (* = hydrophobic; upper/lower case = totally/highly conserved)

N-terminal extensions can also be quite large, particularly in the kexin family (Fig. 2, Table 4). These extensions are quite unique because there is no apparent sequence homology in the large N-terminal extensions. They are often highly charged, like the pro-peptides, but their function is unknown.

Substrate specificity and catalysis

Figure 6 shows our schematic representation of the binding region in subtilases, based on 3D binding data of subtilisins (McPhalen and James, 1988; Heinz et al., 1991; Takeuchi et al., 1991a, 1991b) and thermitase (Gros et al., 1989). This binding region can be described as a surface channel or crevice capable of accommodating at least six amino acid residues (P4–P2′) of a polypeptide substrate or inhibitor (pseudo-substrate). Both main-chain and side-chain interactions between enzyme and substrate/inhibitor contribute to binding. The P4–P1 backbone is H-bonded to the enzyme backbone β-sheet residues 100–102 (strand eI in Fig. 1) and 125–127 (strand eIII), forming the central strand (eII) of a three-stranded antiparallel β-sheet. The C-terminal or leaving portion P1′–P2′ of the substrate appears to be held less tightly as it runs along the enzyme backbone segment 217–219.

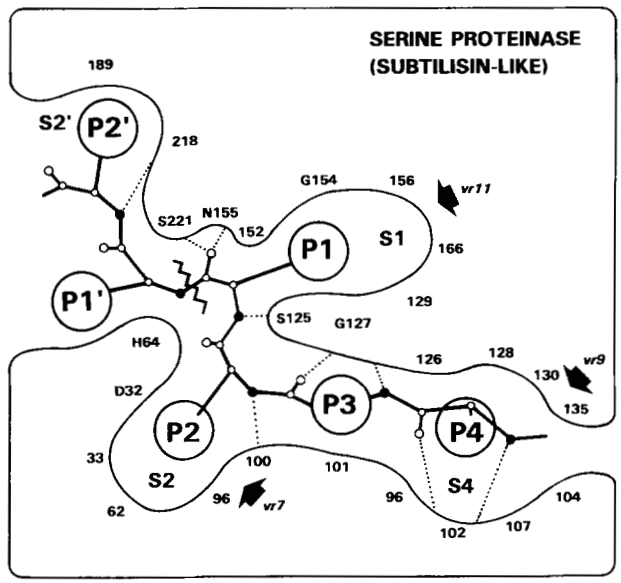

Fig. 6. Schematic representation of substrate/inhibitor (bold lines) binding to a subtilisin-like serine proteinase (smooth surfaces). Nomenclature P4–P2′ and S4–S2′ according to Schechter and Berger (1967). Side chains of the P4–P2′ residues are shown as large spheres; positions of the enzyme residues that may interact with these P4–P2′ side chains are shown surrounding the binding sites (S1, S2, etc). Enzyme numbering is that of subtilisin BPN′. Hydrogen bonds between enzyme and substrate/inhibitor are shown as dotted lines, and the scissile bond is shown by a jagged line. Catalytic residues D32, H64, and S221, and oxyanion-hole residue N155 are indicated. Approximate positions of inserts (vr7, vr9, and vr11) are shown by large arrows.

In general, the specificity of subtilases appears to be largely determined by interactions of the P4–P1 residue side chains in the enzymes' S4–S1 binding sites, respectively, with S4 and S1 dominating the substrate preference in subtilisin (Gron et al., 1992). These sites have the following general characteristics:

S2′: A hydrophobic pocket of variable size depending on the orientation of the conserved aromatic side chain of residue 189.

S1: A distinct, large and elongated cleft, surrounded at the sides and bottom by the backbone segments 125–128 and 152–155, at the bottom end by residue 166 and at the rim by residues 156 and 129.

S2: A less distinct, smaller cleft, bounded at either side by residue 100 and active site residue H64, at the bottom by hydrophobic residue 96 and active site residue D32, at the bottom end by residue 33 and at the rim by residue 62.

S3: Not a distinct site, because the P3 residue points away from the enzyme towards the solvent. However, the most likely interaction is with enzyme residue 101, which is adjacent and points in the same direction. P3 side chains could also interact with nearby residue 100 and the more distant residue 129.

S4: A very distinct pocket, between the segments 101–104 and 126–130, which appears to have two subsites; these subsites can have different characteristics. Site 4a has at the side and bottom the residues 96, 107, and 126, and at the rim 102. Site 4b has at the side and bottom residues 104 and 135, and at the rim residues 128 and 130. These side chains determine the size of the S4

pocket; in subtilisin Y104 is thought to form a flexible lid to the S4 pocket (Takeuchi et al., 1991a).

In subtilisin and thermitase the S1 and S4 binding sites are large and hydrophobic, which explains the broad specificity of both enzymes with a preference for aromatic or large nonpolar P1 and P4 substrate residues (Gron et al., 1992). For further details on the structural basis of substrate specificity in subtilases we refer to the recent review by Perona and Craik (1995).

Variations in the substrate specificity of naturally occurring subtilases should be due to (and could be modified by) modulation of the residues in the substrate-binding region as shown in Figure 6, and in particular those residues whose side chains interact with P1 and P4 substrate residues. Some general predictions of substrate specificity can be made by comparison of the multiple sequence alignment in Figure 2 with the substrate-binding model in Figure 6. Most of the subtilases should have a broad specificity and can be considered as general-purpose proteases, because their binding regions resemble those of subtilisin and thermitase.

The most notable exception occurs when residue 166 at the bottom of the S1 pocket is an Asp, making the protease specific for cleavage after P1 Arg residues. This occurs in family C (ylxpr2), family D (seepip and llnisp), and in all of the kexin family E members; these highly specific proteases are all involved in activation of pro-proteins. A certain preference for cleavage after P1 Lys residues is observed in proteases with a negative charge on residue 156 at the rim of the S1 pocket (Wells et al., 1987; W. Voorhorst, A. Warner, W. de Vos, R. Siezen, in prep.). The kexin family proteases are even more specific because they cleave only after dibasic or multiple basic residues (Barr, 1991; van de Ven et al., 1993). Modeling and engineering studies indicate that a high density of negative charge at the substrate-binding face, and in particular at the S1, S2, and S4 sites, is responsible for this high selectivity (Van de Ven et al., 1990; Creemers et al., 1993; Siezen et al., 1994; Lipkind et al., 1995; Perona & Craik, 1995). These modeling studies predict that the (semi-)conserved acidic residues at positions 33, 61, 97, 104, 107, 129, 130, 131, 161, 166, 191, and 209 in family E subtilases are all in or near the substrate-binding region. Based on this modeling, acidic residues were introduced in the S1 and S2 sites of subtilisin, and this led to a specificity for dibasic residues (Ballinger et al., 1995).

Details of other enzyme–substrate interactions that could be important for substrate binding and selectivity can be obtained by homology modeling, as demonstrated for the family D members efcyla (Booth et al., 1996) and llnisp (Siezen et al., 1995a), the family E members (Siezen et al., 1994), and the family F members llprtp (Siezen et al., 1993), tsplst and pfpyro (W. Voorhorst, A. Warner, W. de Vos, R. Siezen, in prep.). These studies all suggest that electrostatic interactions between enzyme and substrate are more dominant than hydrophobic interactions, and that they are used to generate a more narrow specificity for certain substrates. In addition, interactions with P3, P5, and P6 residues can also contribute to substrate binding, particularly if these are charge–charge interactions. These predictions have been verified by protein engineering of residues involved in enzyme–substrate interactions in llprtp (Siezen et al., 1993), llnisp (Van der Meer et al., 1994; Siezen et al., 1995a), and hsfur (Creemers et al., 1993; Siezen et al., 1994).

Calcium coordination sites

Four calcium-ion binding sites are known from crystal structures of subtilisins, thermitase, and proteinase K; these calcium ions are essential for stability and activity. From previous sequence alignments and homology modeling it was predicted that the Ca1 (strong) and Ca3 (weak) sites are most common in members of the subtilase family, whereas the Ca2 (medium-strength) site is less common (Siezen et al., 1991). The weak Ca4 site has only been found in proteinase K (Betzel et al., 1988a, 1988b).

For the new subdivision into six families the following predictions can be made about the Ca1 and Ca2 sites. The Ca1 site requires coordination from side-chain ligands of residues 2 and 41 and from several side chains of residues 76–81 in the Ca^{2+}-ion embracing loop; these ligands are usually carboxyl/carbonyl groups of Asp/Asn, but can also be from Glu/Gln. This Ca1 site is therefore predicted to be present in nearly all members of families A, B, and E, because they appear to contain the required ligands. In contrast, this Ca1 site cannot be present in any member of families C and D due to the lack of loop 76–81, nor is it likely to occur in family F members due to the high variability in sequence in this loop.

The Ca2 site requires coordination from several side-chain and main-chain ligands of the loop 49–58, with side chains of residues 49, 52, and 54 appearing to be essential, and stabilization by the positively charged side chain Arg/Lys of residue 94. Many family B and E members have the elements required for this Ca2 site if the side-chain oxygen ligands of Asp, Asn, Glu, Gln, Ser, and Thr are considered as acceptable. Some members of families A (intracellular proteases) and C (vaproa, alapr2) should also have the Ca2 site. Predictions for the families D and F are too difficult because in general the sequence alignment is rather speculative in this region; however, some likely candidates for the Ca2 site are seepip, llnisp, bsvpr, llprtp, and ldprtb.

The Ca3 and Ca4 sites are weak and characteristically only have one or two side-chain ligands in the known structures. For this reason no predictions are attempted for these sites in other proteases.

Disulfide bonds

Disulfide bridges can contribute to the overall stability of a protein, and the introduction of new S–S bonds can enhance the thermal stability, as demonstrated in, for example, phage T4 lysozyme (Matsumura et al., 1989). Initial attempts to stabilize subtilisin by introduction of S–S bonds 22–87, 24–87, 26–232, 29–119, 36–210, 41–80, and 148–243 were not successful (Wells & Powers, 1986; Pantoliano et al., 1987; Mitchinson & Wells, 1989; Katz & Kossiakof, 1990); all of these crosslinks were designed by inspection of the three-dimensional structure of subtilisin. The first successful thermal stabilization of subtilisin was the introduction of the 61–98 S–S bond (Takagi et al., 1990), which occurs naturally in aqualysin. It stands to reason, therefore, that naturally occurring S–S bonds should provide a better choice for stabilization of subtilisins than previously designed disulfides.

Seven naturally occurring disulfide bonds have now been identified in subtilases, i.e., 27–118[−2] and 175–247 in proteinase K (taprok; Betzel et al., 1988a, 1988b), 61–98 and 163–195 in aqualysin (taaqua; Kwon et al., 1988), 53–100 and 171–131[+1] in *Dichelobacter* basic protease (dnbpr; Lilley et al., 1992), and 47–59[−1] in *Bacillus* subtilisin S41 (bsta41; Davail et al., 1994). Based on sequence homology (Fig. 2), we predict that these seven S–S bonds also occur in many other subtilases (Table 6). Based on the known three-dimensional structures, together with the sequence alignment in Fig. 2, we also predict that many Cys residues are correctly positioned to form other natural S–S bonds, as listed in Table 6.

Table 6. *Putative and known (in bold) S–S bonds in subtilases*

Family	S–S bond	Context	Enzyme
A	29–114	e1-hD	vmvapt, psaprp
	35–69	Strand-hC	bsispq, tsiap
	47–59[–1]	e2-loop	paalys, bsta39, **bsta41**
	49–55	e2-loop	bssepr
	vr16	Within insert	vmvapt
B	53–100	Turn-turn	dnbpr, xcproa, dnapv2, dnapv5, alapr1
	171–131[+1]	Loop-loop	dnbpr, xcproa, dnapv2, dnapv5, alapr1
	259–263	Intraloop	xcproa, alapr1
C	**27–118[–2]**	e1-hD	**taprok**, tapror, taprot, bbpr1, fusalp, plbspr, macdpa
	175–247	e6-hG	**taprok**, tapror, taprot, bbpr1, fusalp, plbspr, macdpa
	61–98	Loop-turn	**taaqua**, vaproa, alapr2, trt41a
	163–195	loop-turn	**taaqua**, vaproa, alapr2, trt41a, anpepc, spsepr, scprb1, scysp3, ylxpr2
	120–117[+1]	Intraloop	scyct5
	68–224	hC-hF	aoespr
E	80–214	Loop-e9	All except asaspa, avprca
	163–193	Loop-loop	All except asaspa, avprca
	68–224	hC-hF	avprca
	198–254	e7-hG	avprca
	vr1	Within insert	hspac4, hspc6
	vr16	Within insert	asaspa
	vr19	Within insert	cepc2
F	96–102	Intraloop	atserp, cmcucu, lep69, paaf70
	135–167	hE-loop	hstpp2
	151–224	e5-hF	atserp, lep69
	193–197	Intraloop	smserp, phssa1, smssp1, smssp2
	214–75[+3]	e9-loop	hskiaa
	vr4	Within insert	llsp09, cmcucu, agserp, atserp, lep69, paaf70
	vr5	Within insert	hstpp2, cetpp, dmpga9
	vr13	Within insert	llsp09, cmcucu, agserp, atserp, lep69, paaf70, ddtagb, ddtagc
	vr19	Within insert	slssp

Figure 7 shows a stereo view of these known and putative natural S–S bonds, but only those that can be superimposed on the subtilisin BPN′ structure. Other putative S–S bonds may occur in large inserts that have more than one Cys residue (Table 6). Of the 17 "natural" S–S bonds shown in Figure 7, only 29–114 and 163–193 were included in a theoretical prediction of the 31 most energetically and stereochemically favorable disulfide conformations in subtilisin (Hazes & Dijkstra, 1988). Two natural S–S bonds

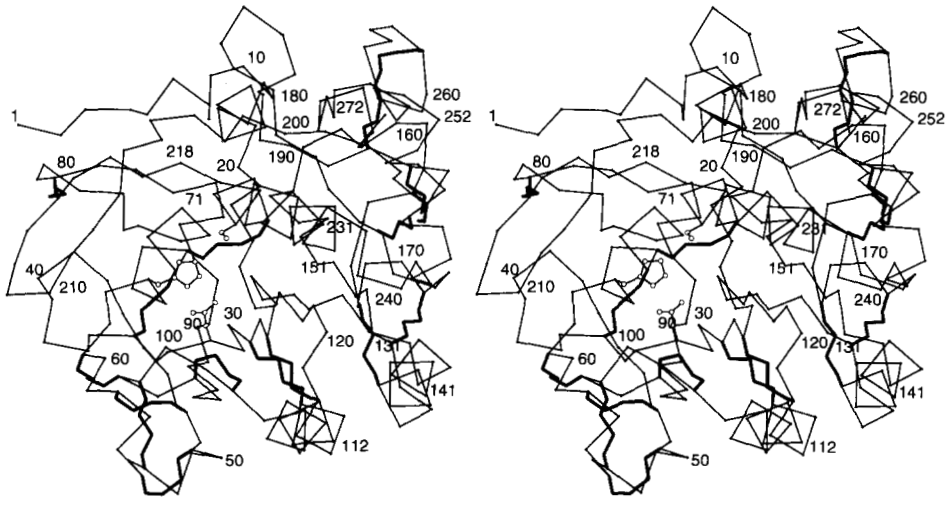

Fig. 7. Stereo view of known and putative natural disulfide bonds (bold) in subtilase members, superimposed on the subtilisin BPN′ structure (α-carbon atom trace). Side chains of the catalytic residues are shown in ball-and-stick representation.

appears to be the maximum for any subtilase, with the possible exception of atserp (Table 6). All the remaining Cys residues in Figure 2 should not form S–S bonds, because they are either single or predicted to be too far removed spatially from another Cys residue.

Disulfides are predicted in each of the families A to F, except for family D, the lantibiotic leader peptidases. These disulfides appear to be family specific, and in some cases even sub-family specific. Extracellular enzymes of gram-positive bacteria rarely contain disulfides. While Cys residues are indeed rare in extracellular subtilases from gram-positive bacteria (Fig. 2), a disulfide has been found in a subtilisin (bsta41) from a psychophilic *Bacillus*, and a different disulfide is predicted in another extracellular subtilase (bssepr) from *Bacillus* (Table 6).

Conclusion

New members of the subtilase superfamily are being identified continuously, with even more to be expected from the accelerating genome sequencing projects. Therefore, this summary is bound to be incomplete when it appears in print. The fact that subtilases have now been discovered in numerous Archaea, Bacteria, and Eucarya suggests that they are ubiquitous and have been around for a long time. The novel information accumulated since our previous review (Siezen et al., 1991) provides exciting new insights into this unique set of enzymes. Through evolution, many variants have arisen and at present these can be divided into six main families A to F (Fig. 4), based on sequence alignment of only the catalytic domains. This classification is by no means definitive yet, as a further subdivision of family F may become apparent when more sequences are available. Subtilases are quite common in gram-positive bacteria, and *Bacillus* species stand out in particular, as many extracellular and even intracellular variants have been identified (Tables 1 and 2), belonging to four different families. Recently, a *Bacillus* strain was even found to have a cluster of four different subtilase genes (Schmidt et al., 1995), belonging to families A and F.

What is most surprising now is the high degree of sequence variability that is observed within the catalytic domains of subtilases. With the exception of the three catalytic residues Asp-His-Ser virtually every other residue can be replaced by one or more different residues. Moreover, it is not even clear what the minimal structural framework requirement is, because large deletions have now been found (Table 5). Large insertions in this domain are also quite common (Table 5), and it is still not clear whether these additions provide extra stability, binding sites, or other functionalities. In one case, Bruinenberg et al. (1994a) demonstrated that deletion of such a large insert (151 residues) in *Lactococcus lactis* proteinase PrtP did not impair protein folding, but it did affect proteolytic activity and specificity. While sequence comparisons and homology modeling can provide a first estimate of overall structure and functionality, and are useful as a tool for rational design of engineered enzymes (Siezen et al., 1991), the high sequence and structural variabilities observed here clearly make some of the predictions speculative and emphasize the need for more detailed 3D structural information to complement the sequence data.

High sequence variability is also found in other protease families, such as in the trypsin family of serine proteases (Rypniewski et al., 1994) and the papain family of cysteine proteases (Berti & Storer, 1995), although these both tend to have more conserved disulfides. Subtilases do not rely on highly conserved disulfides for stabilization, and in fact, most subtilases do not have any disulfides. When these enzymes do have disulfides there is presumably a maximum of two bonds, which can occur in many different positions (Table 6).

Known members of the subtilase superfamily are all (putative) endoproteases or tripeptidylpeptidases. In most bacteria, archaea, and lower eukaryotes they are extracellular, rather unspecific enzymes required either for defense or for growth on proteinaceous substrates. In certain cases, and particularly in higher eukaryotes, the subtilases have developed into highly specialized enzymes of biosynthetic pathways where they are involved in processing and maturation of pro-proteins; examples are all family D and E members. As yet, no other completely different function appears to have arisen for this protein through evolution. On the other hand, given the high sequence variability allowed for proteases, it is questionable whether such a protein would be recognized as a subtilase superfamily member in database screening if it has also lost one or more of the three conserved catalytic residues. Nevertheless, the search is still on, and the authors would appreciate any useful comments, updates or advice.

Acknowledgments

We are greatly indebted to our many colleagues for communicating their sequence data prior to publication. We thank Drs. S. Visser and O. Kuipers for critically reading this manuscript. Use of the services and facilities of the Dutch National NWO/SURF Expertise Center CAOS/CAMM is gratefully acknowledged.

References

Abraham LD, Breuil C. 1995. Factors affecting autolysis of a subtilisin-like serine proteinase secreted by *Ophiostoma piceae* and identification of the cleavage site. *Biochim Biophys Acta 1245*:76–84.

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

Balaban NP, Sharipova MR, Itskovich EL, Leshchinskaya IB, Rudenskaya GN. 1994. Secreted serine protease from the spore-forming bacterium *Bacillus intermedius* 3–19. *Biochemistry (Moscow) 59*:1033–1038.

Ballinger MD, Tom J, Wells JA. 1995. Designing subtilisin BPN' to cleave substrates containing dibasic residues. *Biochemistry 34*:13312–13319.

Barr PJ. 1991. Mammalian subtilisins: The long-sought dibasic processing endoproteases. *Cell 66*:1–3.

Barrett AJ, Rawlings ND. 1995. Families and clans of serine peptidases. *Arch Biochem Biophys 318*:247–250.

Benjannet S, Lusson L, Hamelin J, Savaria D, Chrétien M, Seidah NG. 1995. Structure–function studies on the biosynthesis and bioactivity of the precursor convertase PC2 and the formation of the PC2/7B2 complex. *FEBS Lett 362*:151–155.

Berti PJ, Storer AC. 1995. Alignment/phylogeny of the papain superfamily of cysteine proteases. *J Mol Biol 246*:273–283.

Betzel C, Belleman M, Pal GP, Bajorath J, Saenger W, Wilson KS. 1988a. X-ray and model-building studies on the specificity of the active site of proteinase K. *Proteins Struct Funct Genet 4*:157–164.

Betzel C, Pal GP, Saenger W. 1988b. Three-dimensional structure of proteinase K at 0.15 nm resolution. *Eur J Biochem 178*:155–171.

Booth MC, Bogie ChP, Sahl HG, Siezen RJ, Hatter KL, Gilmore MS. 1996. Structural analysis and proteolytic activation of *Enterococcus faecalis* cytolysin, a novel lantibiotic. *Mol Microbiol 21*:1175–1184.

Bruinenberg PG, Doesburg P, Alting AC, Exterkate FA, Vos WM de, Siezen RJ. 1994a. Evidence for a large dispensable segment in the subtilisin-like catalytic domain of the *Lactococcus lactis* cell-envelope proteinase. *Protein Eng 7*:991–996.

Bruinenberg PG, Vos WM de, Siezen RJ. 1994b. Prevention of C-terminal autoprocessing of *Lactococcus lactis* SK11 cell-envelope proteinase by engineering of an essential surface loop. *Biochem J 302*:957–963.

Burton KS, Wood DA, Thurston CF, Barker PJ. 1993. Purification and characterization of a serine proteinase from senescent sporophores of the commercial mushroom *Agaricus bisporus*. *J Gen Microbiol 139*:1379–1386.

Carter P, Wells JA. 1990. Functional interaction among catalytic residues in subtilisin BPN'. *Proteins Struct Function Genet 7*:335–342.

Chestukhina GG, Zagnit'ko OP, Revina LP, Klepikova FS, Stepanov VM. 1986. Extracellular serine proteinases from subspecies of *Bacillus thuringiensis* evolve much more slowly than the corresponding δ-endotoxins. *Biochemistry (Moscow) 51*:1472–1479.

Creemers JWM, Siezen RJ, Roebroek AJM, Ayoubi TAY, Heylebroeck D, Van de Ven WJM. 1993. Modulation of furin-mediated proprotein processing activity by site-directed mutagenesis. *J Biol Chem 268*:21826–21834.

Davail S, Feller G, Narinx E, Gerday C. 1994. Cold adaptation of proteins. *J Biol Chem 269*:17448–17453.

Devereux J, Haeberli P, Smithies O. 1984. A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res 12*:387–395.

Durham DR. 1993. The elastolytic properties of subtilisin GX from alkalophilic *Bacillus* sp. strain 6644 provides a means of differentiation from other subtilisins. *Biochem Biophys Res Commun 194*:1365–1370.

Felsenstein J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Seattle, Department of Genetics, University of Washington.

Freeman SA, Peek K, Prescott M, Daniel R. 1993. Characterization of a chelator-resistant proteinase from *Thermus* strain Rt4A2. *Biochem J 295*:463–369.

Gallagher T, Bryan P, Gilliland GL. 1993. Calcium-independent subtilisin by design. *Proteins 16*:205–213.

Gallagher T, Gilliland G, Wang L, Bryan P. 1995. The prosegment-subtilisin BPN' complex: Crystal structure of a specific 'foldase.' *Structure 3*:907–914.

Gaucher GM, Stevenson KJ. 1976. Thermomycolin. *Methods Enzymol 45*:415–433.

Greer J. 1990. Comparative modeling methods: Application to the family of mammalian serine proteases. *Proteins Struct Funct Genet 7*:317–334.

Gron H, Meldal M, Breddam K. 1992. Extensive comparison of the substrate preferences of two subtilisins as determined with peptide substrates which are based on the principle of intramolecular quenching. *Biochemistry 31*:6011–6018.

Gros P, Betzel Ch, Dauter Z, Wilson KS, Hol WGJ. 1989. Molecular dynamics refinement of a thermitase-eglin-c complex at 1.98 Å resolution and comparison of two crystal forms that differ in calcium content. *J Mol Biol 210*:347–367.

Hazes B, Dijkstra BW. 1988. Model building of disulfide bonds in proteins with known three-dimensional structure. *Protein Eng 2*:119–125.

Heinz DW, Priestle JP, Rahuel J, Wilson KS, Grütter MG. 1991. Refined crystal structures of subtilisin Novo in complex with wild-type and two mutant eglins: Comparison with other serine proteinase inhibitor complexes. *J Mol Biol 217*:353–371.

Heringa J, Argos P, Egmond MR, Vlieg J de. 1995. Increasing thermal stability of subtilisin from mutations suggested by strongly interacting side-chain clusters. *Protein Eng 8*:21–30.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, ed. *Mammalian protein metabolism*, vol. III. New York: Academic Press. pp 21–132.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*:2577–2637.

Katz B, Kossiakoff AA. 1990. Crystal structures of subtilisin BPN' variants containing disulfide bonds and cavities: Concerted structural rearrangements induced by mutagenesis. *Proteins Struct Funct Genet 7*:343–357.

Kimura M. 1983. *The neutral theory of molecular evolution*. Cambridge, UK: Cambridge University Press.

Kraulis PJ. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structure. *J Appl Crystal 24*:946–950.

Kunitate A, Okamoto M, Ohmori I. 1989. Purification and characterization of a thermostable serine protease from *Bacillus thuringiensis*. *Agric Biol Chem 53*:3251–3256.

Kwon ST, Terada I, Matsuzawa H, Ohta T. 1988. Nucleotide sequence of the gene for aqualysin I (a thermophilic alkaline serine protease) of *Thermus aquaticus* YT-1 and characteristics of the deduced primary structure of the enzyme. *Eur J Biochem 173*:491–497.

Kwon YT, Kim JO, Moon SY, Lee HH, Rho HM. 1994. Extracellular alkaline proteases from alkalophilic *Vibrio metschnikovii* strain RH530. *Biotech Lett 16*:413–418.

Larcher G, Cimon B, Symoens F, Tronchin G, Charbasse D, Bouchara J-P. 1996. A 33 kDa serine proteinase from *Scedosporium apiospermum*. *Biochem J 315*:119–126.

Lavrenova GI, Gul'nik SV, Kalugar SV, Borovikova VP, Revina LP, Stepanov VM. 1984. Extracellular acid serine proteinase D of *Streptomyces rutgersensis*. *Biochemistry (Moscow) 49*:447–454.

Lilley G, Stewart DJ, Kortt AA. 1992. Amino acid and DNA sequences of an extracellular basic protease of *Dichelobacter nodosus* show that it is a member of the subtilisin family of proteases. *Eur J Biochem 210*:13–21.

Lipkind G, Gong Q, Steiner DF. 1995. Molecular modeling of the substrate specificity of prohormone convertases SPC2 and SPC3. *J Biol Chem 270*:13277–13284.

Lo RYC, Strathdee CA, Shewen PE, Cooney BJ. 1991. Molecular studies of SsaI, a serotype-specific antigen of *Pasteurella haemolytica* A1. *Infect Immun 59*:3398–3406.

Matsumura M, Signor G, Matthews BW. 1989. Substantial increase of protein stability by multiple disulphide bonds. *Nature 342*:291–293.

McPhalen CA, James MNG. 1988. Structural comparison of two serine proteinase–protein inhibitor complexes: Eglin-C-subtilisin Carlsberg and CI-2-subtilisin Novo. *Biochemistry 27*:6582–6598.

Meyer C, Bierbaum G, Heidrich C, Reis M, Süling J, Iglesias-Wind MI, Kempter C, Molitor E, Sahl HG. 1995. Nucleotide sequence analysis of the lantibiotic gene cluster and functional analysis of PepP and PepC. *Eur J Biochem 232*:478–489.

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

Murphy JM, Walton JD. 1996. Three extracellular proteases from *Cochliobolus carbonum*: Cloning and targeted disruption of *ALP1*. *Mol Plant-Microbe Interact 9*:290–297.

Pantoliano MW, Ladner RC, Bryan PN, Rollence ML, Wood JF, Poulos TL. 1987. Protein engineering of subtilisin BPN': Enhanced stabilization through the introduction of two cysteines to form a disulfide bond. *Biochemistry 26*:2077–2082.

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA 85*:2444–2448.

Perona JJ, Craik CS. 1995. Structural basis of substrate specificity in the serine proteases. *Protein Sci 4*:337–360.

Rawlings ND, Barrett AJ. 1994. Families of serine peptidases. *Methods Enzymol 244*:19–61.

Rypniewski WR, Perrakis A, Vorgias CE, Wilson KS. 1994. Evolutionary divergence and conservation of trypsin. *Protein Eng 7*:57–64.

Sahl HG, Jack RW, Bierbaum G. 1995. Biosynthesis and biological activities of lantibiotics with unique post-translational modifications. *Eur J Biochem 230*:827–853.

Saitou N, Nei M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol 4*:406–425.

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

Schmidt BF, Woodhouse L, Adams RM, Ward T, Mainzer SE, Lad PJ. 1995. Alkalophilic *Bacillus* sp. strain LG12 has a series of serine protease genes. *Appl Environ Microbiol 61*:4490–4493.

Segers R, Butt TM, Keen JN, Kerry BR, Peberdy JF. 1995. The subtilisins of the invertebrate mycopathogens *Verticillium chlamydosporium* and *Metarhizium anisopliae* are serologically and functionally related. *FEMS Microbiol Lett 126*:227–232.

Shimogaki H, Takeuchi K, Nishino T, Ohdera M, Kudo T, Ohba K, Iwama M, Irie M. 1991. Purification and properties of a novel surface-active agent- and alkaline-resistant protease from *Bacillus* sp. Y. *Agric Biol Chem 55*:2251–2258.

Siezen RJ, Vos WM de, Leunissen JAM, Dijkstra BW. 1991. Homology modelling and protein engineering strategy of subtilases, the family of subtilisin-like serine proteases. *Protein Eng 4*:719–737.

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

Siezen RJ, Creemers JWM, van de Ven WJM. 1994. Homology modelling of the catalytic domain of human furin. A model for the eukaryotic subtilisin-like proprotein convertases. *Eur J Biochem 222*:255–266.

Siezen RJ, Rollema HS, Kuipers OP, Vos WM de. 1995a. Homology modelling of the *Lactococcus lactis* leader peptidase NisP and its interaction with the precursor of the lantibiotic nisin. *Protein Eng 8*:117–125.

Siezen RJ, Leunissen JAM, Shinde U. 1995b. Homology analysis of the propeptides of subtilisin-like serine proteases (subtilases). In: Shinde U, ed. *Intramolecular chaperones and folding*. Austin: R.G. Landes Company. pp 231–253.

Strongin AYa, Izotova LS, Abramov ZT, Gorodetsky DI, Ermakova LM, Baratova LA, Belyanova LP, Stepanov VM. 1978. Intracellular serine protease of *Bacillus subtilis*: Sequence homology with extracellular subtilisins. *J Bacteriol 133*:1401–1411.

Takagi H, Takahashi T, Momose H, Inouye M, Maeda Y, Matsuzawa H, Ohta T. 1990. Enhancement of the thermostability of subtilisin E by introduction of a disulfide bond engineered on the basis of structural comparison with a thermophilic serine protease. *J Biol Chem 265*:6874–6878.

Takeuchi Y, Noguchi S, Satow Y, Kojima S, Kumagai I, Miura K, Nakamura KT, Mitsui Y. 1991a. Molecular recognition at the active site of subtilisin BPN': Crystallographic studies using genetically engineered proteina-

ceous inhibitor SSI (*Streptomyces* subtilisin inhibitor). *Protein Eng 4*:501–508.

Takeuchi Y, Satow Y, Nakamura KT, Mitsui Y. 1991b. Refined crystal structure of the complex of subtilisin BPN′ and *Streptomyces* subtilisin inhibitor at 1.8 Å resolution. *J Mol Biol 221*:309–325.

Tsujibo H, Miyamoto K, Hasegawa T, Inamori Y. 1990. Amino acid compositions and partial sequences of two types of alkaline serine proteases from *Nocardiopsis dassonvillei* subsp. *prasina* OPC-210. *Agric Biol Chem 54*:2177–2179.

van de Ven WJM, Voorberg J, Fontijn R, Pannekoek H, Ouweland AMW van den, Duijnhoven HLP van, Roebroek AJM, Siezen RJ. 1990 Furin is a subtilisin-like proprotein processing enzyme in higher eukaryotes. *Mol Biol Rep 14*:265–275.

van de Ven WJM, Roebroek AJM, Duijnhoven HLP van. 1993. Structure and function of eukaryotic proprotein processing enzymes of the subtilisin family of serine proteases. *Crit Rev Oncogenesis 4*:115–136.

van der Meer JR, Rollema HS, Siezen RJ, Kuipers OP, Vos WM de. 1994. Influence of amino acid substitutions in the nisin leader peptide on biosynthesis and secretion of nisin by *Lactococcus lactis*. *J Biol Chem 269*:3555–3562.

Wells JA, Powers DB. 1986. *In vivo* formation and stability of engineered disulfide bonds in subtilisin. *J Biol Chem 261*:6564–6570.

Wells JA, Powers DB, Bott RR, Graycar TP, Estell DA. 1987. Designing substrate specificity by protein engineering of electrostatic interactions. *Proc Natl Acad Sci USA 84*:1219–1223.

Zhou A, Paquet L, Mains E. 1995. Structural elements that direct specific processing of different mammalian subtilisin-like prohormone convertases. *J Biol Chem 270*:21509–21516.