Biochem. J. (1993) 290, 205–218 (Printed in Great Britain)

Evolutionary families of peptidases

Neil D. RAWLINGS and Alan J. BARRETT

Department of Biochemistry, Strangeways Research Laboratory, Worts Causeway, Cambridge CB1 4RN, U.K.

The available amino acid sequences of peptidases have been examined, and the enzymes have been allocated to evolutionary families. Some of the families can be grouped together in 'clans' that show signs of distant relationship, but nevertheless, it appears that there may be as many as 60 evolutionary lines of peptidases with separate origins. Some of these contain members with quite diverse peptidase activities, and yet there are some striking examples of convergence. We suggest that the classification by families could be used as an extension of the current classification by catalytic type.

INTRODUCTION

Amino acid sequence data are now available for over 600 peptidases (endopeptidases, exopeptidases and omega peptidases), and we have examined these in an attempt to establish what separate evolutionary lines exist. These take the form of families, or groups of related families ('clans'). The properties of the peptidases of each family have been considered from two main points of view. Firstly, we have asked how widely the enzymes have *diverged* in catalytic activity, and, secondly, we have asked to what extent peptidases from separate evolutionary lines have *converged* in properties. Finally, we have considered how compatible is a classification of peptidases based on their evolutionary relationships with the sort of classification that is currently in use, which depends upon the reaction catalysed by each enzyme and on the catalytic mechanism.

METHODS

Sources of data

Protein sequence data were obtained from the SwissProt database [1] (release 21), and the PIR-Protein database [2] (release 32), and nucleic acid sequence data from the EMBL database [1] (release 28 and daily updates). In addition, some sequences were obtained directly from the literature.

Detection of evolutionary relationships

The programs FASTP [3] and FASTA and TFASTA [4] were used to detect similarities between peptidases, and, on the basis of these, provisional assignments to a system of families was made. These assignments were refined by manual construction of optimized alignments. In many cases, the similarities between the sequences were so close that no further analysis was felt necessary, but whenever the similarity was questionable, the RDF program [3] was applied. This tests the statistical significance of a similarity between amino acid sequences by comparing the score for the alignment with those of random shuffles of the sequences. We took the value of six standard deviation units as that above which the similarity could be regarded as being significant. We assume that the significant similarities reflect evolutionary relationship, or homology as defined by Reeck et al. [5].

Definition of terms

The term *type* is used to refer to a set of peptidases distinguished according to the chemical groups responsible for catalysis, as in serine-type, cysteine-type, aspartic-type or metallo-type. The term *family* is used to describe a group of enzymes in which each member shows evolutionary relationship to at least one other, either throughout the whole sequence or at least in the part of the sequence responsible for catalytic activity. As an example of the need for this, bone morphogenetic protein 1 is a chimaeric protein that contains a catalytic domain related to that of astacin, but also contains segments that are clearly homologous with non-catalytic parts of C1r and C1s in the chymotrypsin family [6]. We place bone morphogenetic protein 1 in the family of astacin and not in that of chymotrypsin.

A *clan* comprises a group of families for which there are indications of evolutionary relationship, despite the lack of statistically significant similarities in sequence. Such indications of distant relationship come primarily from the linear order of catalytic-site residues and the tertiary structure. Distinctive aspects of the catalytic activity such as specificity or inhibitor-sensitivity may also contribute occasionally.

The symbol '+' is used to indicate the scissile bond in a peptidase substrate.

RESULTS AND DISCUSSION

All of the amino acid sequences of peptidases that were available to us in July 1992 were examined for significant similarities as described in the Methods section, and grouped in families (Table 1). Some of the families show evidence of distant relationships to others, and these we group together in single 'clans'; others seem quite unrelated.

Serine peptidases

Most of the members of the chymotrypsin (S1) family are endopeptidases, which differ widely in specificity. No exopeptidase is known in this family, but it does contain several proteins that lack all peptidase activity: azurocidin, procarboxypeptidase A complex component III, the haptoglobins, apolipoprotein a, hepatocyte growth factor and protein Z. The family includes many enzymes of the coagulation, fibrinolysis and complement systems that are found in blood plasma, and these are mostly chimaeric proteins with modules, some of which are also found in other proteins, inserted N-terminally to the site of proteolytic activation [27].

Almost all of the known members of the chymotrypsin family have been found in animals, the only exceptions being two trypsins from actinomycetes. It is striking that no member of this otherwise very successful family has been encountered in protozoa, fungi or plants.

The linear order of catalytic triad residues in the polypeptide

Table 1 Evolutionary families of peptidases

The peptidases are allocated to families as described in the text. Clans and families are labelled with the prefix S for serine peptidases, C for cysteine, A for aspartic, M for metallo- and U for unknown, and listed in this order. It should be noted, however, that these labels are temporary, simply being assigned consecutively through the Table. 'EC' is the enzyme nomenclature number [7], but for peptidases the initial '3.4.' has been omitted; '-' indicates that no EC number has been assigned; 'n.a.' indicates that the protein is not known to be an enzyme. Literature references to the individual proteins are generally to be found in the database entries for which the codes are given. Most of the codes are from the Swiss-Prot database (release 21), but a code in parentheses is an EMBL database accession number and 'PIR' indicates a code from the PIR database. Numbers in square brackets are references to sequences from journal articles. For some viral sequences, the code given is that of the viral polyprotein. For some viruses, numerous variants with only minor differences exist, and only a single example of each has been included.

	EC	Database code
SERINE PEPTIDASES		
Family S1: Chymotrypsin	*(Clan SA: His, Asp, Ser catalytic triad)*	
Trypsin (includes forms I, II, III, IV Va and Vb)	21.4	TRYP_SACER, TRYP_STRGR, TRYP_ASTFL, TRYP_DROME, TRYP_SQUAC, TRYP_XENLA, TRYP_BOVIN, TRY1_CANFA, TRY2_CANFA, TRY1_HUMAN, TRY2_HUMAN, TRY3_HUMAN, TRYP_MOUSE, TRYP_PIG, TRY1_RAT, TRY2_RAT, TRY3_RAT, TRY4_RAT, (M77814), (X59012), (X59013)
Cercarial elastase (*Schistosoma*)	-	CERC_SCHMA
Brachyurin	21.32	COGS_UCAPU
Factor C (*Limulus*)	-	(D90271)
Proclotting enzyme (*Tachypleus*)	-	PCE_TACTR
easter gene product (*Drosophila*)	-	EAST_DROME
snake gene product (*Drosophila*)	-	SNAK_DROME
Vitellin-degrading endopeptidase *Bombyx*)	-	[8]
Hypodermin C	21.49	COGS_HYPLI
Serine proteases 1 and 2 (*Drosophila*)	-	SER1_DROME
Achelase (*Lonomia*)	-	ACH1_LONAC, ACH2_LONAC
Chymotrypsin (includes forms A, B, II and 2)	21.1	CTR2_VESCR, CTR2_VESOR, CTRA_BOVIN, CTRB_BOVIN, CTR2_CANFA, CTRB_HUMAN, CTRB_RAT
Proteinase RVV-V (Russell's viper) (includes forms α and γ)	-	RVVA_VIPRU, RVVG_VIPRU
Flavoboxin (habu snake)	-	FLVB_TRIFL
Venombin A	21.74	BATX_BOTAT, PTCA_AGKCO
Crotalase	21.74	[9]
Enteropeptidase	21.9	[10]
Acrosin	21.10	ACRO_HUMAN, ACRO_MOUSE, ACRO_PIG
Seminin	-	PROS_HUMAN
Tissue kallikrein	21.35	KAG2_CAVPO, KAG1_HUMAN, KAG2_HUMAN, KAG_PIG, KAGP_RAT
Renal kallikrein	21.35	KAGR_MOUSE, (X17352)
Submandibular kallikrein	21.35	KAG1_MOUSE, KAG2_MOUSE, KAG3_MOUSE, KAG5_MOUSE, KAGB_MOUSE, KAG1_RAT, KAG3_RAT
7S nerve growth factor (includes α and γ chains)	21.35	NGFA_MOUSE, NGFG_MOUSE
Epidermal growth factor-binding protein (includes forms 1, 2 and 3)	21.35	EGBA_MOUSE, EGBB_MOUSE, EGBC_MOUSE
Tonin	21.35	TONI_RAT
Arginine esterase	21.35	ESTA_CANFA
Pancreatic elastase I	21.36	EL1_PIG, EL1_RAT, (M27347)
Pancreatic elastase II (includes forms A and B)	21.71	EL2A_HUMAN, EL2B_HUMAN, EL2_MOUSE, EL2_PIG, EL2_RAT
Pancreatic endopeptidase E (includes forms A and B)	21.70	EL3A_HUMAN, EL3B_HUMAN
Leukocyte elastase	21.37	ELNE_HUMAN
Medullasin	-	ELNE_HUMAN
Azurocidin	n.a.	CAP7_HUMAN, CAP7_PIG

Table 1 (*contd.*)

Cathepsin G	21.20	CATG_HUMAN
Proteinase 3 (myeloblastin)	-	MELB_HUMAN, PTN3_HUMAN
Chymase (includes forms I and II)	21.39	MCP1_CANFA, TRYM_CANFA, MCP1_MOUSE, MCP2_MOUSE, MCP1_RAT, MCP2_RAT, MCP4_MOUSE, (M69136), (M73759)
γ-Renin	21.54	RENG_MOUSE
Tryptase (includes forms 1, 2 and 3)	21.59	TRYT_CANFA, TRYA_HUMAN, TRYB_HUMAN, (M33493), (M30038), MCP6_MOUSE
Hepsin	-	HEPS_HUMAN
Granzyme A	-	GRAA_HUMAN, GRAA_MOUSE, GRAX_MOUSE
Natural killer cell protease 1	-	NKP1_RAT
Granzymes B, C, D, E, F, G and Y	-	GRAB_MOUSE, GRAC_MOUSE, GRAD_MOUSE, GRAE_MOUSE, GRAF_MOUSE, GRAG_MOUSE, GRAB_HUMAN, GRAY_HUMAN
Carboxypeptidase A complex component III	n.a.	CAC3_BOVIN
Complement factor D	21.46	CFAD_HUMAN, ADIP_MOUSE
Complement factor B	21.47	CFAB_HUMAN, CFAB_MOUSE
Complement factor I	21.45	CFAI_HUMAN
Complement component C$\overline{1r}$	21.41	CO1R_HUMAN
Complement component C$\overline{1s}$	21.42	C1S_HUMAN
Calcium-dependent serine proteinase	-	CASP_MESAU
Complement component C2	21.43	CO2_HUMAN, CO2_MOUSE
Haptoglobin (includes forms 1 and 2)	n.a.	HPT1_HUMAN, HPT2_HUMAN
Haptoglobin-related protein	n.a.	HPTR_HUMAN
Plasmin	21.7	PLMN_BOVIN, PLMN_HUMAN, PLMN_MACMU, PLMN_MOUSE, PLMN_PIG, (M62832)
Apolipoprotein(a)	n.a.	APOA_HUMAN, APOA_MACMU
Hepatocyte growth factor	n.a.	HGF_HUMAN, HGF_RAT
Thrombin	21.5	THRB_BOVIN, THRB_HUMAN, THRB_MOUSE, THRB_RAT
t-Plasminogen activator	21.68	UROT_HUMAN, UROT_MOUSE, UROT_RAT
u-Plasminogen activator	21.73	UROK_CHICK, UROK_HUMAN, UROK_MOUSE, UROK_PAPCY, UROK_PIG
Salivary plasminogen activator (vampire bat)	21.68	UROT_DESRO
Plasma kallikrein	21.34	KAL_HUMAN, KAL_RAT, (M58588)
Coagulation factor VII	21.21	FA7_BOVIN, FA7_HUMAN
Coagulation factor IX	21.22	FA9_BOVIN, FA9_CANFA, FA9_HUMAN, FA9_MOUSE
Coagulation factor X	21.6	FA10_BOVIN, FA10_HUMAN
Coagulation factor XI	21.27	FA11_HUMAN
Coagulation factor XII	21.38	FA12_HUMAN
Protein C	21.69	PRTC_BOVIN, PRTC_HUMAN
Protein Z	n.a.	PTRZ_BOVIN, PRTZ_HUMAN
Family S2: α-Lytic endopeptidase	(*Clan SA: His, Asp, Ser catalytic triad*)	
α-Lytic endopeptidase	21.12	PRLA_LYSEN
Proteases A and B (*Streptomyces griseus*)	-	PRTA_STRGR, PRTB_STRGR
Glutamyl endopeptidase (*Strep. griseus*)	-	[11]
Family S3: Togavirus endopeptidase	(*Clan SA: His, Asp, Ser catalytic triad*)	
Polyprotein peptidase	-	POLS_EEEV, POLS_RRVN, POLS_SFV, POLS_SINDV, POLS_WEEV
Family S4: Glutamyl endopeptidase		
Glutamyl endopeptidase (*Staphylococcus*)	21.19	STSP_STAAU
Epidermolytic toxins A and B (*Staphylococcus*)	-	ETA_STAAU, ETB_STAAU
"Metalloprotease" (*Bacillus subtilis*)	-	[12]
Family S5: Lysyl endopeptidase		
Lysyl endopeptidase (*Achromobacter*)	21.50	API_ACHLY
Family S6: IgA-specific endopeptidase		
IgA-specific serine endopeptidase	21.72	IGA_NEIGO, (X64357)

Table 1 (*contd.*)

Family S7: Flavivirus endopeptidase		
Nonstructural protein NS3	-	POLG_DEN2J, POLG_JAEVJ, POLG_KUNJM, POLG_MVEV, POLG_TBEVS, POLG_WNV, POLG_YEFV1
Family S8: Subtilisin		(*Asp, His, Ser catalytic triad*)
Tripeptidyl-peptidase II	14.10	(M73047)
Subtilisin	21.62	SUBT_BACAM, SUBT_BACLI, SUBT_BACMS, SUBT_BACSA, SUBT_BACSD, SUBT_BACSU
Alkaline elastase (*Bacillus*)	-	ELYA_BACSU
Serine endopeptidase (*Bac. subtilis*)	-	(PIR S11504)
Major intracellular endopeptidase (*Bacillus*)	-	ISP1_BACSU, (D00862), (D10730)
Bacillopeptidase F (*Bac. subtilis*)	-	SUBF_BACSU
Neutral endopeptidase (*Bacillus*)	-	NPRE_BACAM, NPRE_BACSU
Thermitase	21.66	THET_THEVU
C5a peptidase (*Streptococcus*)	-	SCPA_STRPY
Cell-wall associated endopeptidase (*Lactococcus*) (forms PI, PII, PIII)	-	P1P_LACLA, P2P_LACLA, P3P_LACLA
Aqualysin I (*Thermus*)	-	AQL1_THEAQ
Extracellular endopeptidase (*Serratia*)	-	PRTS_SERMA
Calcium-dependent extracellular endopeptidase A (*Vibrio*)	-	PROA_VIBAL
Extracellular endopeptidase (*Xanthomonas*)	-	PIR S11890
Endopeptidase K	21.64	PRTK_TRIAL
Endopeptidase R (*Tritirachium*)	-	PRTR_TRIAL
Endopeptidase T (*Tritirachium*)	-	PRTT_TRIAL
Cuticle-degrading protease (*Metarhizium*)	-	(M73795)
Oryzin	21.63	AEP_ASPOR, AEP_YARLI
Alkaline protease (*Aspergillus*)	-	(Z11580)
Cerevisin	21.48	PRTB_YEAST
Subtilisin-like protease III (*Saccharomyces*)	-	(M77197)
Alkaline endopeptidase (*Acremonium*)	-	PIR JU0332
Calcium dependent endopeptidase (*Anabaena*)	-	PRCA_ANAVA
Kexin	21.61	KEX2_YEAST, KEX1_KLULA
Furin	-	FURI_HUMAN, FURI_MOUSE, FURI_RAT, (M81431)
Pituitary convertase (includes PC1 and PC2)	-	NEC1_MOUSE, NEC2_HUMAN, NEC2_MOUSE
Family S9: Prolyl oligopeptidase		(*Asp, Ser, His or Ser, Asp, His catalytic triad*)
Dipeptidyl-peptidase IV	14.5	DPP_RAT, (X60708)
Dipeptidyl aminopeptidase B (*Saccharomyces*)	-	DAP2_YEAST
Acylaminoacyl-peptidase	19.1	ACPH_PIG, ACPH_RAT
Protease II (*Escherichia coli*)	-	TLP_ECOLI
Prolyl oligopeptidase	21.26	PPCE_PIG, (M81461), (M61966)
DNF1552 protein (3p21 protein)	n.a.	DNF1_HUMAN
Family S10: Serine-type carboxypeptidase		(*Ser, Asp, His catalytic triad*)
Serine-type carboxypeptidase (*Saccharomyces*)	16.1	CBPY_YEAST, (D10199)
Carboxypeptidase B-like peptidase	16.1	KEX1_YEAST, CBP2_HORVU, CBP2_WHEAT,
Serine-type carboxypeptidase (forms I and III)	16.1	CBP1_HORVU, CBP3_HORVU, CBP3_WHEAT, (D10985)
Carboxypeptidase Y-like protein (*Arabidopsis*)	-	(M81130)
Serine-type carboxypeptidase (*Caenorhabditis*)	-	(M75784)
Serine-type carboxypeptidase (*Aedes*)	-	(M79452)
Lysosomal carboxypeptidase A	16.1	PRTP_HUMAN, PRTP_MOUSE

Table 1 (*contd.*)

Family S11: D-Ala-D-Ala carboxypeptidase (gene *daca*) (*Clan SB: Ser, Lys, Ser, Glu catalytic tetrad*)		
Serine-type D-Ala-D-Ala carboxypeptidase	16.4	DACA_BACSU, DACA_ECOLI, DACC_ECOLI, (X59965), (M37688)
Family S12: D-Ala-D-Ala carboxypeptidase (gene *dac*) (*Clan SB: Ser, Lys, Ser, Glu catalytic tetrad*)		
Serine-type D-Ala-D-Ala carboxypeptidase	16.4	DAC_STRSP
D-Aminopeptidase (*Ochrobactrum*)	-	(M84523)
ß-lactamase	3.5.2.6	AMPC_CITFR, AMPC_ECOLI, AMPC_ENTCL, AMPC_SERMA
Protein FIMD (*Bacteroides*)	-	FMDH_BACNO, FMDD_BACNO
Family S13: Penicillin-binding protein 4 (*Clan SE: Ser, Lys, Ser, Glu catalytic tetrad*)		
Serine-type D-Ala-D-Ala carboxypeptidase	16.4	[13]
Penicillin-binding protein 4	16.4	PBP4_ECOLI
Family S14: ClpP (*Ser, His catalytic residues (Asp not known)*)		
ATP-dependent endopeptidase (ClpP subunit) (*Escherichia coli*)	-	CLPP_ECOLI
Chloroplast ATP-dependent endopeptidase	-	CLPP_MARPO, CLPP_TOBAC, CLPP_ORYSA, CLPP_WHEAT
Potato leaf roll luteovirus genomic RNA	n.a.	(D00530), (X14600)
Family S15: *Lactococcus* dipeptidyl peptidase IV		
Dipeptidyl peptidase IV (*Lactococcus*)	14.5	DPP_LACLA, DPP_LACLC
Family S16: Endopeptidase La		
Endopeptidase La	21.53	LON_ECOLI, (D00863)
Family S17: *Bacteroides* endopeptidase		
Extracellular endopeptidase (*Bacteroides*)	-	PRTE_BACNO
Family S18: Endopeptidase VII		
Protease VII (*Escherichia coli*)	-	OMPT_ECOLI
Coagulase/fibrinolysin (*Yersinia*)	-	COLY_YERPE
Phosphoglycerate transport system activator (*Salmonella*)	-	PGTE_SALTY
Family S19: *Coccidioides* endopeptidase		
Chymotrypsin-like protease (*Coccidioides*)	-	(X63114)
Family S20: Protease Do		
Protease Do (*Salmonella*)	-	(X54548)
Family S21: Assemblin, herpesvirus		
Assemblin	-	UL26_HSV11, VG33_VZVD, CP40_ILV, YEC3_EBV, UL80_HCMVA, (M64627)
Family S22: Placental protein 11		
Placental protein 11	-	PP11_HUMAN
CYSTEINE PEPTIDASES		
Family C1: Papain (*Clan CA: Gln, Cys, His, Asn active site residues*)		
Dipeptidyl peptidase I	14.1	(D90404)
Cysteine endopeptidases 1 (*Haemonchus*)	-	CYS1_HAECO,
Cysteine endopeptidases 1 (*Haemonchus*)	-	(M80385)
Surface protective protein (*Plasmodium*)	n.a.	[14]
Circumsporozoite protein (*Plasmodium*)	-	CSP_PLACM
Cysteine endopeptidase (*Entamoeba*)	-	(M27307), (M64712), (M64721)
Cysteine endopeptidase (*Trypanosoma*)	-	CYSP_TRYBR
Cruzipain (*Trypanosoma*)	-	(M90067)
Cysteine endopeptidase (*Theileria*)	-	CYSP_THEPA, (M86659)
Cysteine endopeptidase (*Leishmania*)	-	(X62163)
Cysteine endopeptidases 1 and 2 (*Dictyostelium*)	-	CYS1_DICDI, CYS2_DICDI
Endopeptidase (baculovirus of *Autographa*)	-	(M67451)
Papain	22.2	PAPA_CARPA
Chymopapain	22.6	PAP2_CARPA
Caricain	22.30	PAP3_CARPA

Table 1 (*contd.*)

Glycyl endopeptidase	22.25	PAP4_CARPA
Actinidain	22.14	ACTN_ACTCH
Cysteine endopeptidase (tomato)	-	CYSL_LYCES
Thaumatopain (*Thaumatococcus*)	-	THPA_THADA
Calotropin (*Calotropis*)	-	CAL1_CALGI
Cysteine endopeptidase (*Brassica napus*)	-	[15]
Cysteine endopeptidase (mung bean)	-	SHEP_VIGMU
Endopeptidase EP-C1 (*Phaseolus vulgaris*)	-	(X63102)
Protein P34 (soya bean)	n.a.	P34_SOYBN
Clone 15a protein (garden pea)	-	[16]
Stem bromelain	22.32	BROM_ANACO
Aleurain (barley)	-	ALEU_HORVU
Cysteine endopeptidases 2 and 3 (barley)	-	[17]
Oryzain (includes forms α, β and γ) (rice)	-	[18]
Cysteine protease (*Caenorhabditis*)	-	(M74797)
Cysteine endopeptidases 1, 2 and 3 (*Homarus*)	-	(X63567), (X63568), (X63569)
Allergen (*Dermatophagoides*)	-	MMAL_DERPT
Allergen (*Euroglyphus*)	-	(X60073)
Cathepsin L	22.15	CATL_CHICK, CATL_HUMAN, CATL_MOUSE, CATL_RAT
Cathepsin S	22.27	CATS_BOVIN, (M86553)
Cathepsin H	22.16	CATH_HUMAN, CATH_RAT
Cathepsin B	22.1	CATB_BOVIN, CATB_HUMAN, CATB_MOUSE, CATB_RAT, (M75822), (M21309)
Family C2: Calpain	(*Clan CA: Gln, Cys, His, Asn active site residues*)	
Sol gene product (*Drosophila*)	-	(M64084)
Calpain (*Schistosoma*)	22.17	(M67499)
Calpain I	22.17	CAP1_CHICK, CAP1_HUMAN, CAP1_RABIT
Calpain II	22.17	CAP2_HUMAN, CAP2_RABIT
Calpain P94	22.17	CAP3_HUMAN, CAP3_RAT
Calcium-binding protein PMP41	22.17	CAP4_MOUSE
Family C3: Picornain	(*Clan CB: His, Asp or Glu, Cys catalytic triad*)	
Picornain 2A	22.29	POLG_POL1M, POLG_COXA2, POLG_SVDVH, POLG_BOVEV, POLG_HRV14
Picornain 3C	22.28	POLH_POL1M, POLG_COXA2, POLG_SVDVH, POLG_BOVEV, POLG_HRV14, POLG_ECHO9, POLG_TMEVD
Aphthovirus endopeptidase		POLG_FMDVD
Cardiovirus endopeptidase		POLG_EMCV
Comovirus endopeptidase		VGNB_CPMV, (D00657)
Family C4: Potyvirus endopeptidase 1	(*Clan CB: His, Asp, Cys catalytic triad*)	
48 kDa endopeptidase	-	POLG_PPVD, POLG_PPVRA, POLG_PPVYN, POLG_TEV, POLG_TVMV, POLG_WMV2, POLG_OMV
Family C5: Adenovirus endopeptidase	(*Clan CB: His, Cys catalytic triad*)	
Endopeptidase adenovirus	-	VPRT_ADEB3, VPRT_ADEB7, VPRT_ADE02, VPRT_ADE03, VPRT_ADE04, VPRT_ADE05, VPRT_ADE12, VPRT_ADE40, VPRT_ADE41, (M81056)
Family C6: Potyvirus endopeptidase 2		
29 kDa endopeptidase	-	POLG_PPVD, POLG_PVYN, POLG_TEV, POLG_TVMV
Family C7: Chestnut blight virus p29 endopeptidase		
p29 Endopeptidase (Chestnut blight virus)	-	(M57938)
Family C8: Chestnut blight virus p48 endopeptidase		
p48 Endopeptidase (Chestnut blight virus)	-	(M57938)
Family C9: Togavirus cysteine endopeptidase		
Togavirus cysteine endopeptidase	-	POLN_SINDV, POLN_RRVN, POLN_SFV,POLN_ONNVG,

Table 1 (*contd.*)

Family C10: Streptopain		
Streptopain	22.10	STCP_STRPY
Family C11: Clostripain		
α-Clostripain	22.8	CLOL_CLOHI
Family C12: Ubiquitin hydrolase		
Ubiquitin carboxyl-terminal hydrolase	-	UBL1_HUMAN, UBL3_HUMAN, [19]
Family C13: Haemoglobinase		
Haemoglobinase (*Schistosoma*)	-	HGLB_SCHMA
Family C14: Interleukin-1β converting enzyme		
Interleukin-1β converting enzyme	-	[20]
ASPARTIC PEPTIDASES		
Family A1: Pepsin	*(Clan AA: Asp, Asp catalytic residues)*	
Aspergillopepsin I	23.18	PEPA_ASPAW
Penicillopepsin	23.20	PENP_PENJA
Rhizopuspepsin	23.21	CARP_RHICH, CARP_RHINI,
Endothiapepsin	23.22	CARP_CRYPA
Mucorpepsin	23.23	CARP_RHIMI, CARP_RHIPU
Candidapepsin	23.24	CARP_CANAL, (X61438), (Z11918), (M83663), (X56867), (Z11919)
Polyporopepsin	23.29	CARP_IRPLA
Saccharopepsin	23.25	CARP_SACFI, CARP_YEAST, (D10198)
"Barrier" protein (*Saccharomyces*)	-	BAR1_YEAST
Aspartic proteinase (barley)	-	(X56136)
Pepsin A	23.1	PEPA_CHICK, PEPA_BOVIN, PEPA_HUMAN, PEPA_MACFU, PEPA_MACMU, PEPA_PIG,
Aspartic endopeptidase P111	-	PIR JT0398
Gastricsin	23.3	PEPC_HUMAN, PEPC_MACFU, PEPC_RAT
Chymosin	23.4	CHYM_BOVIN, CHYM_SHEEP
Embryonic pepsin (chicken)	-	PEPE_CHICK
Renin, submandibular	23.15	RENS_MOUSE
Renin, renal	23.15	RENI_HUMAN, RENI_MOUSE, RENI_RAT
Cathepsin D	23.5	CATD_HUMAN, CATD_MOUSE, CATD_PIG, CATD_RAT
Cathepsin E	23.34	CATE_HUMAN
Family A2: Retropepsin	*(Clan AA: Asp, Asp catalytic residues)*	
Retropepsin	23.16	POL_HIV1A, POL_HIV2D, POL_SIVMK, POL_BIV06, POL_EIAV, POL_VILV, VPRT_MPMV, VPRT_MMTVB, GAG_RSVP, VPRT_BLV, POL_FLV, POL_GALV, VPRT_HTL1A, POL_MLVAV, VPRT_SMRVH, VPRT_SRV1
Retrovirus-related endopeptidase (human)	-	VPRT_HUMAN
Retropepsin-like protein (vaccinia virus)	-	(M25392)
METALLO-PEPTIDASES		
Family M1: Alanyl aminopeptidase	*(Clan MA: Peptidases with HEXXH zinc-binding motif)*	
Membrane alanyl aminopeptidase	11.2	AMPN_ECOLI, AMPN_HUMAN, AMPN_PIG, AMPN_RAT, (X51508), (M75750)
Lysyl aminopeptidase (*Lactococcus*)	11.15	(X61230)
Aminopeptidase yscII (*Saccharomyces*)	-	(X63998)
BP-1/6C3 antigen, mouse	-	BP1_MOUSE
Leukotriene A_4 hydrolase	3.3.2.6	LKHA_HUMAN, (M63848)
Family M2: Peptidyl-dipeptidase A	*(Clan MA: Peptidases with HEXXH zinc-binding motif)*	
Peptidyl-dipeptidase A	15.1	ACE_HUMAN, ACET_HUMAN, ACE_MOUSE, ACET_MOUSE, ACE_RABIT, ACET_RABIT

Table 1 (*contd.*)

Family M3: Thimet oligopeptidase		(*Clan MA: Peptidases with HEXXH zinc-binding motif*)
Peptidyl-dipeptidase, bacterial	-	(X57947), (M84575)
Oligopeptidase (*Salmonella*)	-	(M84574)
Mitochondrial intermediate peptidase	-	(M96633)
Saccharolysin	24.37	(X59720 - orf YCL57w)
Thimet oligopeptidase	24.15	MEPD_RAT
Family M4: Thermolysin		(*Clan MA: Peptidases with HEXXH zinc-binding motif*)
Thermolysin	24.27	THER_BACST, THER_BACTH
Pseudolysin	24.26	ELAS_PSEAE
Neutral endopeptidase (*Bacillus stearothermophilus*)	-	PIR B36706
Bacillolysin	24.28	THER_BACCE, THER_BACCL, NPRE_BACSU, (D00861), (K02497), (M64815), (X61380)
Metalloendopeptidase (*Legionella*)	-	PROA_LEGPN
Vibriolysin (*Vibrio*)	-	(M64809), (M59466)
Extracellular endopeptidase (*Erwinia*)	-	(M36651)
Metalloendopeptidase (*Listeria*)	-	PROL_LISMO
Coccolysin	24.30	(M37185)
Family M5: Mycolysin		(*Clan MA: Peptidases with HEXXH zinc-binding motif*)
Mycolysin	24.31	NPR_STRCI
Family M6: Immune inhibitor A		(*Clan MA: Peptidases with HEXXH zinc-binding motif*)
Immune inhibitor A (*Bacillus thuringiensis*)	-	INA_BACTL
Family M7: *Streptomyces* small neutral protease		(*Clan MA: Peptidases with HEXXH zinc-binding motif*)
Small neutral protease (*Streptomyces*)	-	(M81703), (M86606), (Z11929)
Family M8: Leishmanolysin		(*Clan MA: Peptidases with HEXXH zinc-binding motif*)
Leishmanolysin	24.36	GP63_LEICH, GP63_LEIDO, GP63_LEIMA, (X64394)
Family M9: Microbial collagenase		(*Clan MA: Peptidases with HEXXH zinc-binding motif*)
Collagenase (*Vibrio*)	24.3	[21]
Family M10: Interstitial collagenase		(*Clan MA: Peptidases with HEXXH zinc-binding motif*)
Serralysin	24.40	PRTB_ERWCH, PRTC_ERWCH, PRTX_ERWCH, PRZN_SERSP
Envelysin	24.12	HE_PARLI
Matrilysin	24.23	COG7_HUMAN
Interstitial collagenase	24.7	COG1_HUMAN, COG1_PIG, COG1_RABIT
Neutrophil collagenase	24.34	COG8_HUMAN
Stromelysin 1	24.17	COG3_HUMAN, COG3_RABIT, COG3_RAT
Stromelysin 2	24.22	COGX_HUMAN, COGX_RAT
Stromelysin 3	-	COGY_HUMAN
Gelatinase A	24.24	GOG2_HUMAN
Gelatinase B	24.35	COG9_HUMAN
Family M11: Autolysin		(*Clan MA: Peptidases with HEXXH zinc-binding motif*)
Autolysin	24.38	[22]
Family M12: Astacin		(*Clan MA: Peptidases with HEXXH zinc-binding motif*)
Metalloendopeptidase (*Caenorhabditis*)	-	(M75746)
Blastula protease-10 (*Paracentrotus*)	-	(X56224)
Astacin	24.21	ASTA_ASTFL
tolloid gene product (*Drosophila*)	-	(M76976)
UVS.2 protein (*Xenopus*)	-	[23]
Ruberlysin	24.48	HRT2_CRORU
Atrolysin c	24.42	HRTD_CROAT
Trimerelysin II	24.53	HR2_TRIFL
HR2a-endopeptidase (habu snake)	-	HR2A_TRIFL
HR1B-endopeptidase (habu snake)	-	HR1B_TRIFL
Haemorrhagic factor LHFII (bushmaster snake)	-	HRL2_LACMU
Meprin A	24.18	(M74897)

Table 1 (*contd.*)

PABA-peptide hydrolase	24.18	(M82962)
Bone morphogenetic protein 1	-	BMP1_HUMAN
Family M13: Neprilysin	(*Clan MA: Peptidases with HEXXH zinc-binding motif*)	
Neprilysin	24.11	NEP_HUMAN, NEP_RABIT, NEP_RAT
Kell blood group protein	-	KELL_HUMAN
Family M14: Carboxypeptidase A	(*HXXE zinc-binding motif*)	
Zinc-carboxypeptidase (*Streptomyces*)	-	CBPS_STRGR
Carboxypeptidase T (*Thermoactinomyces*)	-	(X56901)
Carboxypeptidase B	17.2	CBPB_ASTFL, CBPB_BOVIN, CBPB_RAT, (M75106)
Carboxypeptidase A	17.1	CBPA_BOVIN, CBPC_HUMAN, CBPC_MOUSE, CBP1_RAT, CBP2_RAT, (A25833)
Lysine carboxypeptidase	17.3	CBPN_HUMAN
Carboxypeptidase H	17.10	CBPH_BOVIN, CBPH_HUMAN, CBPH_RAT, (X61232), [24]
Carboxypeptidase M	17.12	CBPM_HUMAN
Family M15: Muramoyl-pentapeptide carboxypeptidase	(*HXH zinc-binding motif*)	
Muramoyl-pentapeptide carboxypeptidase	17.8	CBPM_STRGR
Family M16: Pitrilysin	(*HXXEH zinc-binding motif*)	
Pitrilysin	99.44	PTR_ECOLI
pqqF gene product (*Klebsiella*)	-	(X58778)
Insulinase	99.45	IDE_DROME, IDE_HUMAN
Mitochondrial processing peptidase	99.41	MPP1_NEUCR, MPP1_YEAST, MPP1_RAT
Processing enhancing protein	-	MPP2_NEUCR, MPP2_YEAST
Ubiquinol-cytochrome c reductase core proteins 1 and 2	1.6.99.3	UCR1_YEAST, UCR2_YEAST, UCR2_HUMAN
Family M17: Leucyl aminopeptidase	(*Peptidases binding two zinc atoms: Lys, Glu, Asp, Asp, Glu*)	
Leucyl aminopeptidase	11.1	AMPL_BOVIN, (X63444)
Aminopeptidase A (*Escherichia coli*)	-	AMPA_ECOLI, (M68966)
Family M18: Aminopeptidase yscI		
Aminopeptidase yscI (*Saccharomyces*)	-	AMPL_YEAST, LAP4_YEAST
Family M19: Membrane dipeptidase		
Membrane dipeptidase	13.19	MDP4_HUMAN, MDP4_PIG
Open reading frame X product (*Klebsiella*)	-	(X58778)
Gene R product (*Acinetobacter*)	-	(X06452)
Family M20: Carboxypeptidase G2		
Carboxypeptidase G2 (*Pseudomonas*)	-	CBPG_PSES6
Peptidase T (*Salmonella*)	-	(M62725)
Family M21: Gly-X carboxypeptidase		
Gly-X carboxypeptidase (*Saccharomyces*)	17.4	(X57316)
Family M22: A1 Glycoprotease		
A1 Glycoprotease (*Pasteurella*)	-	(M62364)
OrfX (*Escherichia coli*)	-	YRUX_ECOLI
OrfX (*Salmonella*)	-	(M14427)
Family M23: ß-lytic endopeptidase		
ß-Lytic endopeptidase	24.32	PRLB_LYSEN, (M60896)
LasA protein (*Pseudomonas*)	-	LASA_PSEAE
Family M24: Methionyl aminopeptidase		
Methionyl aminopeptidase	11.18	AMPM_BACSU, AMPM_ECOLI, AMPM_SALTY
Aminopeptidase P (*Escherichia coli*)	-	AMPP_ECOLI
X-Pro dipeptidase	13.9	PEPQ_ECOLI, PEPD_HUMAN
Family M25: X-His-dipeptidase		
X-His dipeptidase	13.3	PEPD_ECOLI
PEPTIDASES OF UNKNOWN CATALYTIC TYPE		
Family U1: Aminopeptidase T		
Aminopeptidase T (*Thermus*)	-	AMPT_THEAQ

Table 1 (*contd.*)

Family U2: Aminopeptidase IAP		
Alkaline phosphatase isozyme conversion protein (*Escherichia coli*)	-	IAP_ECOLI
Family U3: Spore endopeptidase, *Bacillus megaterium*		
Spore endopeptidase (*Bacillus megaterium*)	-	(M55262)
Family U4: Sporulation sigma factor processing peptidase		
Sporulation sigma factor processing peptidase (*Bacillus subtilis*)	-	SP2G_BACSU
Family U5: Tail-specific protease		
Tail-specific protease (*Esherichia coli*)	-	(M75634)
Family U6: Murein endopeptidase		
Penicillin-insensitive murein endopeptidase (*Esherichia coli*)	-	MEPA_ECOLI
Family U7: Endopeptidase IV		
Endopeptidase IV (*Escherichia coli*)	-	SPPA_ECOLI, LICA_HAEIN
sohB gene product (*E. coli*)	-	(M73320)
Minor capsid protein precursor C (bacteriophage lambda)	-	VCAC_LAMBD
Family U8: Bacteriophage endopeptidase		
Endopeptidase (bacteriophage)	-	ENPP_BPPA2, ENPP_BPP22, ENPP_BPT3, ENPP_BPT7, ENPP_LAMBD
Family U9: Prohead endopeptidase		
Prohead endopeptidase (bacteriophage T4)	-	PCPP_BPT4
Family U10: Leader peptidase		
Leader peptidase	99.36	LEP_ECOLI, LEP_SALTY, (X56466), (Z11847)
Mitochondrial inner membrane peptidase 1 (*Saccharomyces*)	-	[25]
Family U11: Premurein leader peptidase		
Premurein leader peptidase	99.35	LPSA_ECOLI, LPSA_ENTAE, LPSA_PSEFL, (M83994), (M84707)
Family U12: Prepilin leader peptidase		
Prepilin leader peptidase (*Vibrio*)	-	(M74708)
Late competence protein (*Bacillus*)	-	COMC_BACSU
xpcA protein (*Pseudomonas*)	-	PILD_PSEAE
Pullulanase secretion protein (*Klebsiella*)	-	PULO_KLEPN
Family U13: Leader peptidase component 3-4		
Leader peptidase 21 kDa subunit (dog)	99.36	SPC3_CANFA
Leader peptidase 18 kD subunit (dog)	99.36	SPC4_CANFA
Leader peptidase (*sec11*) (*Saccharomyces*)	99.36	[26]
Family U14: Leader peptidase component 2		
Leader peptidase 22-23 kDa subunit (dog)	99.36	SPC2_CANFA
Microsomal leader peptidase (chicken)	99.36	(X60795)
Leader peptidase (*Drosophila*)	99.36	(M32022)
Family U15: Multicatalytic endopeptidase complex		
Multicatalytic endopeptidase subunits	99.46	PRCA_THEAC, (M83674), (J05358), PRC1_YEAST, PRC7_YEAST, (M63641), PRCD_YEAST, PRCB_YEAST, PRCX_YEAST, PRCZ_YEAST, PR28_DROME, PR29_DROME, PR35_DROME, PRC3_XENLA, (X62709), PRC2_RAT, PRC5_RAT, PRC3_RAT, PRC8_RAT, PRC9_RAT, (M64992), (D00760), (D00761), (D00762), (D00763), (D10729), (X64449)
SCL1 suppressor protein (*Saccharomyces*)	99.46	SCL1_YEAST
Family U16: Thermopsin		
Thermopsin	99.43	THPS_SULAC
Family U17: Ubiquitin-specific processing protease		
Ubiquitin-specific processing protease I (*Saccharomyces*)	-	(M63484)

Table 1 (*contd.*)

Family U18: Scytalidiapepsin		
Scytalidiapepsin B	23.32	PRTB_SCYLI
Scytalidiapepsin (*Aspergillus*)	-	PRTA_ASPNG
Family U19: Pestivirus endopeptidase		
Endopeptidase (cattle viral diarrhoea virus)	-	(M37795), (M62430)
Family U20: γ-D-Glutamyl-L-diamino acid endopeptidase II		
γ-D-glutamyl-L-diamino acid endopeptidase II (*Bacillus sphaericus*)	-	(X64809)
Family U21: Potyvirus endopeptidase 3		
35 kDa endopeptidase	-	POLG_PPVD, POLG_PPVRA, POLG_PPVYN, POLG_TEV, POLG_TVMV

chains of the enzymically active members of family **S1** is His, Asp, Ser. The same order of residues is seen in family **S2** (α-lytic endopeptidase) and family **S3** (togavirus endopeptidase), and members of these families also have tertiary structures similar to that of chymotrypsin [28,29]. This strongly suggests that they share a common evolutionary origin, despite the differences of sequence, and accordingly we group families **S1**, **S2** and **S3** in a single clan (SA). The evidence is less complete for families **S4**, **S5**, **S6** and **S7**, but there are indications that these also may belong in this clan [30–33].

The enzymes of the subtilisin (**S8**) family have a different order of catalytic-site residues from chymotrypsin, namely Asp, His, Ser, and also have different tertiary structures. It is therefore quite clear that the family represents a separate evolutionary line of serine peptidases [34]. The family contains an exopeptidase (tripeptidyl peptidase II) as well as endopeptidases with various specificities. Most of the microbial members of the family have specificities somewhat like that of chymotrypsin, but the eukaryote enzymes include the proprotein convertases such as kexin and furin, which are specific for substrates containing paired basic residues [35].

We consider that the family of prolyl oligopeptidase (**S9**) reflects a further distinct evolutionary line of serine peptidases. In this family there is again a different order of catalytic residues, Ser^{554} and His^{680} being known for pig prolyl oligopeptidase [36]. We have suggested that if an Asp residue completes a catalytic triad, Asp^{529} is the most likely [37]. There is evidence that prolyl oligopeptidase differs significantly in catalytic mechanism from the enzymes of families **S1** and **S8** [38,39]. The family contains two endopeptidases with the restricted specificity for substrate size that makes them oligopeptidases [37]; one of these cleaves prolyl bonds, whereas the other acts on bonds with a basic residue in the P1 position. The family also contains a dipeptidyl peptidase and an omega peptidase [37].

The serine-type carboxypeptidases form family **S10**, in which the order of catalytic residues is Ser, Asp, His. The tertiary structure of these enzymes is unlike those known for other families, and they are unusual amongst serine-type hydrolases in being maximally active at about pH 5 [40]. There are similarities between the structures of the active sites of these enzymes and those lipases [40] and acetylcholinesterases.

There are three distinct families of serine-type D-Ala-D-Ala carboxypeptidases, **S11**, **S12** and **S13**, all confined to bacteria. Their members are similar in catalytic mechanism and three-dimensional structure [13,41,42], and thus are grouped in a single clan (**SB**), which also contains other penicillin-binding proteins. Family **S12** also contains a D-aminopeptidase, the only serine-type aminopeptidase reported to date.

The Clp endopeptidase is one of the ATP-dependent proteolytic enzymes of *Escherichia coli* and contains two subunits, ClpP and ClpA, ClpP being responsible for the peptidase activity. The active-site serine (Ser^{111}) and histidine (His^{136}) of ClpP are known, but no aspartic acid that might form the third member of a catalytic triad has been identified. Other members of this family (**S14**) occur in plant chloroplasts, which may reflect their endosymbiont origins. We report here that the 5′ end of potato-leaf-roll-luteovirus genomic RNA, which has been described as a non-coding region [43], also is homologous with ClpP.

The active-site serine of endopeptidase La (**S16**) has been determined as Ser^{679} [44], but otherwise no catalytic-site residues have been identified in families **S15**–**S22**. Family **S15** contains a dipeptidyl-peptidase specific for the cleavage of Xaa-Pro + bonds that is unrelated to the enzyme with similar specificity in family **S9**.

Cysteine peptidases

In addition to many endopeptidases, the papain family (**C1**) contains an exopeptidase, dipeptidyl-peptidase I, and proteins that lack peptidase activity, in *Plasmodium* and soya bean. Unusually, the papain family contains a sequence from a baculovirus genome, which may have been acquired from a host [45]. With this exception, proteins of the papain family have been found only in eukaryotes.

The calpains (family **C2**) are heterodimeric enzymes, the larger (80 kDa) subunit containing the proteolytic domain and also calcium-binding, E–F-hand structures similar to those found in other proteins. No active-site histidine has yet been positively identified, but Cys^{108} and His^{265} (in chicken calpain) occur in sequences that show some similarity to those around catalytic residues in the papain family [46], and accordingly the families of papain and calpain form a clan (**CA**). The homologous *sol* protein from *Drosophila* has a distinctive structure, being a much larger protein with zinc fingers but no calcium-binding sites [47].

Families **C3**–**C9** comprise viral enzymes. We group families **C3**–**C5** in a clan (**CB**) in which the order of catalytic residues is His, Cys. It is possible that these enzymes are related to those of the chymotrypsin family (**S1**), with interconversion of the essential serine and cysteine residues [48,49]. Such a relationship would represent the only known homology across catalytic types. In most members of the clan an aspartic residue is thought to form the third member of a catalytic triad, but in the picornains 3C, the essential Asp is replaced by Glu [50,51].

Essential cysteine and histidine residues occur in the order Cys, His in families **C6**–**C10**, but families **C6**–**C14** cannot as yet be assigned to clans.

Aspartic peptidases

At present, it seems that all of the aspartic peptidases are endopeptidases, and the great majority of these are members of the pepsin (**A1**) family, which have been found only in eukaryotic organisms. The peptidases of this family have bilobed molecules resulting from gene duplication [52]. In contrast, the viral retropepsins (family **A2**) have a monomeric structure in which each molecule contains only half of the functional catalytic site and dimerization is needed to form the active enzyme. We place families **A1** and **A2** in a clan (**AA**) on the basis of similarities in tertiary structure [53] and sensitivity to inhibition by pepstatin [54].

The catalytic residues of acid proteinases in two other families have not been identified, so it is not possible at this stage to say whether these pepstatin-resistant enzymes also are aspartic endopeptidases. These are the families of thermopsin (**U16**) and scytalidopepsin (**U18**). For scytalidopepsin B it was suggested that Glu^{53} is involved in catalysis [55], but this is not conserved in the second member of the family [56].

Metallopeptidases

The structures of metallopeptidases are exceptionally diverse, and we recognize 25 families. The majority of the enzymes contain zinc, and for several of them the residues involved in binding the zinc have been identified by X-ray crystallography.

Of the families of metallopeptidases, 13 contain the sequence HEXXH, which is known or suspected to provide two of the three ligands for the zinc atom. These are the families of alanyl aminopeptidase (**M1**), peptidyl-dipeptidase A (**M2**), thimet oligopeptidase (**M3**), thermolysin (**M4**), mycolysin (**M5**), immune inhibitor A (**M6**), *Streptomyces* protease (**M7**), leishmanolysin (**M8**), *Vibrio* collagenase (**M9**), interstitial collagenase (**M10**), autolysin (**M11**), astacin (**M12**), and neprilysin (**M13**). With some reservations, we group these in a clan (**MA**). In each of these families, the HEXXH sequence occurs in a nine-residue consensus sequence, bXHEbbHbc, in which b is an uncharged residue, c is hydrophobic, and X can be any amino acid. The third ligand of the zinc atom is glutamic acid in the families of alanyl aminopeptidase [57], thermolysin [58] and neprilysin [59], but is histidine in the astacin family [60], and presumably also in those of autolysin and interstitial collagenase, since the histidine is conserved. Family **M1** contains aminopeptidases and an ether hydrolase (EC 3.3.2.6), and family **M3** contains peptidyl-dipeptidases and oligopeptidases. The large thermolysin family (**M4**) appears to be confined to bacteria, whereas members of the astacin family (**M12**) have been found only in animals.

Families **M10** and **M12** each contain members of only 200–300 residues (astacin and matrilysin respectively), but also much larger chimaeric proteins. Our inclusion of the snake-venom metalloendopeptidases within the astacin family is based on statistically significant sequence relationships with the endopeptidase domain of human bone morphogenetic factor 1 for ruberlysin and atrolysin C.

Four further families of zinc metallopeptidases exhibit distinctive modes of binding of the metal. In the family of carboxypeptidase A (**M14**) the short-spaced ligands of zinc (in the terminology of Vallee [61]) are histidine and glutamic acid in the sequence HXXE, whereas they are two histidine residues in the sequence HXH in the family of muramoylpentapeptide carboxypeptidase (**M15**). In both families, the third ligand is histidine. Even more distinctive is the set of lysine, two aspartic and two glutamic residues that bind a pair of zinc atoms at the active site in the leucyl aminopeptidase (**M17**) family. The enzymes of the pitrilysin family (**M16**) are thought to bind zinc at an HXXEH sequence [62], and in insulinase the sequence is HFCEH, which may account for the thiol-dependence of the enzyme. In pitrilysin, which is not thiol-dependent, this sequence is HYLEH. Several members of the family lack the HXXEH consensus altogether, and presumably are inactive; these include 'mitochondrial processing peptidase' and the reductase subunits. It has been suggested that the activity formerly attributed to mitochondrial processing peptidase is due to the associated processing-enhancing protein [62]. We report here that the *kpqqF* gene of *Klebsiella pneumoniae* [63] shows homology with members of the pitrilysin family; this sequence, which does contain the HXXEH consensus, seems too dissimilar to that of pitrilysin for the *Klebsiella* protein to be simply the species variant.

The groups responsible for zinc-binding are unknown for the other families of metallopeptidases (**M18–M25**). Family **M24** contains aminopeptidases and a dipeptidase.

Unknown catalytic type

There are 21 families for which the catalytic type remains to be determined. The family of endopeptidase IV (**U7**) contains viral and bacterial enzymes, the only such family we know of, but the virus is a bacteriophage, so presumably may have acquired the gene from a host.

The leader peptidases form the five families **U10–U14**. The bacterial leader peptidases (**U10–U12**) are not homologous with eukaryote microsomal leader peptidases, but the bacterial leader peptidase of family **U10** is related to the eukaryotic mitochondrial leader peptidase, which may reflect the endosymbiont origin of mitochondria. The eukaryotic microsomal leader peptidases are multisubunit proteins, and it is not clear which components are directly responsible for peptidase activity. The component subunits form at least two families, **U13** and **U14**. We have noticed that the 3′-terminal portion of a sequence that includes the sex-specific gene *msP316* from *Drosophila* [64] is homologous with the glycoprotein component 2 of microsomal leader peptidase.

The molecules of multicatalytic endopeptidase complex (**U15**) contain two or more kinds of subunit, which are nevertheless homologous. It has yet to be established which of the three or more distinct peptidase activities of the enzyme are attributable to which subunits [65].

Conclusions

Until recently it has seemed that the vast majority of endopeptidases belonged to just a few evolutionary families, those of chymotrypsin, subtilisin, papain, pepsin and thermolysin [66]. By analogy, it might have been expected that the exopeptidases also would prove to belong to just a few families, which would have been separate from those of the endopeptidases. It would thus have been natural to assume that modern peptidases reflect a small number of independent evolutionary origins, perhaps a dozen or so. However, our analysis of the hundreds of amino acid sequences now available for peptidases points to different conclusions. Using rigorous standards for relatedness, we have had to recognize no fewer than 84 distinct families of peptidases. A number of families show signs of distant relationship to others, and accordingly have been placed in clans. Even so, we have 60 groups of sequences among which we can see no relationship, and there may therefore have been this many separate evolutionary origins of peptidases.

The origins of many of the modern families of peptidases clearly were very early, since members are present in modern prokaryotic micro-organisms. Families that appear in prokaryotes, but also are found in eukaryotic organisms, are those of chymotrypsin (**S1**), subtilisin (**S8**), prolyl oligopeptidase (**S9**), ClpP (**S14**), alanyl aminopeptidase (**M1**), thimet oligopeptidase (**M3**), interstitial collagenase (**M10**), carboxypeptidase A (**M14**), pitrilysin (**M16**), leucyl aminopeptidase (**M17**), methionyl aminopeptidase (**M24**), and multicatalytic endopeptidase complex (**U15**). No examples are known for cysteine- or aspartic-type peptidases. Three large families of endopeptidases have few or no known examples among prokaryotes, but underwent major expansion in the eukaryotes; these are the families of chymotrypsin (**S1**), papain (**C1**) and pepsin (**A1**). Possibly they developed in connection with the acquisition of the capacity for endocytosis by eukaryotic cells, to function in membrane-limited organelles, sometimes at acidic pH. All the known viral peptidases are endopeptidases, and the great majority of these show no relationship to the peptidases of other organisms, although exceptions are listed in families **S3**, **C1**, **A2** and **U7** of Table 1. No viral metallopeptidase has been described.

We have seen striking examples of the amount of divergent evolution that can occur in a family of peptidases. Six of the families contain proteins that are not peptidases (**S1**, **S12**, **S14**, **C1**, **M1** and **M16**). Also, four families contain both exopeptidases and endopeptidases (**S8**, **S9**, **C1** and **M3**). Family **S12** contains both aminopeptidases and carboxypeptidases, and family **M24** contains aminopeptidases and a dipeptidase. In most of the families there are peptidases differing greatly in specificity for amino acids around the scissile bond. Conversely, there has been convergence such that a number of peptidase specificities are exhibited by enzymes of more than one family; examples would be the activities of glutamyl endopeptidase (**S2** and **S4**), Xaa-Pro+dipeptidyl peptidase (**S9** and **S15**), peptidyl-dipeptidase (**M2** and **M3**), carboxypeptidase specific for basic residues (**S10** and **M14**) and D-Ala-D-Ala carboxypeptidases (**S11**, **S12**, **S13** and **M15**).

Having constructed a classification of peptidases that is based on structural and evolutionary relationships, we have naturally considered whether this has anything to add to the currently accepted methods of classification by reaction catalysed and by catalytic mechanism. The evolutionary scheme is clearly not compatible with classification of the enzymes by the reactions they catalyse, since many families contain enzymes with quite different kinds of peptidase activities, and some specificities are found in several families. The evolutionary scheme does fit well within the system of classification by catalytic type, however, as can be seen in Table 1, and it tends to bring together the enzymes that resemble each other most closely in structure and catalytic mechanism. We therefore suggest that it deserves serious consideration for use in future schemes for the classification of these enzymes, as an extension of the classification by catalytic type.

We are grateful to the SERC (Science and Engineering Research Council) Daresbury Laboratory for providing access to the sequence databases.

REFERENCES

1 Kahn, P. and Cameron, G. (1990) Methods Enzymol. **183**, 23–31

2 Barker, W. C., George, D. G. and Hunt, L. T. (1990) Methods Enzymol. **183**, 31–49

3 Lipman, D. J. and Pearson, W. R. (1985) Science **227**, 1435–1441

4 Pearson, W. R. and Lipman, D. J. (1988) Proc. Natl. Acad. Sci. U.S.A. **85**, 2444–2448

5 Reeck, G. R., de Haën, C., Teller, D. C., Doolittle, R. F., Fitch, W. M., Dickerson, R. E., Chambon, P., Mclachlan, A. D., Margoliash, E., Jukes, T. H. and Zuckerkandl, E. (1987) Cell **50**, 667

6 Rawlings, N. D. and Barrett, A. J. (1990) Biochem. J. **266**, 622–624

7 Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (1992) Enzyme Nomenclature 1992, Academic Press, Orlando

8 Ikeda, M., Yaginuma, T., Kobayashi, M. and Yamashita, O. (1991) Comp. Biochem. Physiol. B **99**, 405–411

9 Pirkle, H., Markland, F. S., Theodor, I., Baumgartner, R., Bajwa, S. S. and Kirakossian, H. (1981) Biochem. Biophys. Res. Commun. **99**, 715–721

10 Light, A. and Janska, H. (1991) J. Protein Chem. **10**, 475–480

11 Svendsen, I., Jensen, M. R. and Breddam, K. (1991) FEBS Lett. **292**, 165–167

12 Sloma, A., Rudolph, C. F., Rufo, G. A. J., Sullivan, B. J., Theriault, K. A., Ally, D. and Pero, J. (1990) J. Bacteriol. **172**, 1024–1029

13 Granier, B., Duez, C., Lepage, S., Englebert, S., Dusart, J., Dideberg, O., Van Beeumen, J., Frère, J.-M. and Ghuysen, J.-M. (1992) Biochem. J. **282**, 781–788

14 Li, W.-B., Bzik, D. J., Horii, T. and Inselburg, J. (1989) Mol. Biochem. Parasitol. **33**, 13–26

15 Dietrich, R. A., Maslyar, D. J., Heupel, R. C. and Harada, J. J. (1989) Plant Cell **1**, 73–80

16 Guerrero, F. D., Jones, J. T. and Mullet, J. E. (1990) Plant Mol. Biol. **15**, 11–26

17 Koehler, S. M. and Ho, T.-H. D. (1990) Plant Cell **2**, 769–783

18 Watanabe, H., Abe, K., Emori, Y., Hosoyama, H. and Arai, S. (1991) J. Biol. Chem. **266**, 16897–16902

19 Miller, H. I., Henzel, W. J., Ridgeway, J. B., Kuang, W.-J., Chisholm, V. and Liu, C.-C. (1991) Biotechnology **7**, 698–704

20 Thornberry, N. A., Bull, H. G., Calaycay, J. R., Chapman, K. T., Howard, A. D., Kostura, M. J., Miller, D. K., Molineaux, S. M., Weidner, J. R., Aunins, J. et al. (1992) Nature (London) **356**, 768–774

21 Takeuchi, H., Shibano, Y., Morihara, K., Fukushima, J., Inami, S., Keil, B., Gilles, A.-M., Kawamoto, S. and Okuda, K. (1992) Biochem. J. **281**, 703–708

22 Kinoshita, T., Fukuzawa, H., Shimada, T., Saito, T. and Matsuda, Y. (1992) Proc. Natl. Acad. Sci. U.S.A. **89**, 4693–4697

23 Sato, S. M. and Sargent, T. D. (1990) Dev. Biol. **137**, 135–141

24 Roth, W. W., Mackin, R. B., Spiess, J., Goodman, R. H. and Noe, B. D. (1991) Mol. Cell. Endocrinol. **78**, 171–178

25 Behrens, M., Michaelis, G. and Pratje, E. (1991) Mol. Gen. Genet. **228**, 167–176

26 Böhni, P. C., Deshaies, R. J. and Schekman, R. W. (1988) J. Cell Biol. **106**, 1035–1042

27 Patthy, L. (1990) Semin. Thromb. Hemostasis **16**, 245–259

28 Delbaere, L. T. J., Hutcheon, W. L. B., James, M. N. G. and Thiessen, W. E. (1975) Nature (London) **257**, 758–763

29 Choi, H.-K., Tong, L., Minor, W., Dumas, P., Boege, U., Rossmann, M. G. and Wengler, G. (1991) Nature (London) **354**, 37–43

30 Drapeau, G. R. (1978) Can. J. Biochem. **56**, 534–544

31 Bachovchin, W. W., Plaut, A. G., Flentke, G. R., Lynch, M. and Kettner, C. A. (1990) J. Biol. Chem. **265**, 3738–3743

32 Tsunasawa, S., Masaki, T., Hirose, M., Soejima, M. and Sakiyama, F. (1989) J. Biol. Chem. **264**, 3832–3839

33 Chambers, T. J., Weir, R. C., Grakoui, A., McCourt, D. W., Bazan, J. F., Fletterick, R. J. and Rice, C. M. (1990) Proc. Natl. Acad. Sci. U.S.A. **87**, 8898–8902

34 Hartley, B. S. (1970) Philos. Trans. R. Soc. London Ser B **257**, 77–87

35 Barr, P. J. (1991) Cell **66**, 1–3

36 Stone, S. R., Rennex, D., Wikstrom, P., Shaw, E. and Hofsteenge, J. (1991) Biochem. J. **276**, 837–840

37 Barrett, A. J. and Rawlings, N. D. (1992) Biol. Chem. Hoppe-Seyler **373**, 353–360

38 Polgár, L. (1992) Biochem. J. **283**, 647–648

39 Stone, S. R., Rennex, D., Wikstrom, P., Shaw, E. and Hofsteenge, J. (1992) Biochem. J. **283**, 871–876

40 Liao, D.-I. and Remington, S. J. (1990) J. Biol. Chem. **265**, 6528–6531

41 Joris, B., Ghuysen, J.-M., Dive, G., Renard, A., Dideberg, O., Charlier, P., Frère, J.-M., Kelly, J. A., Boyington, J. C., Moews, P. C. and Knox, J. R. (1988) Biochem. J. **250**, 313–324

42 Kelly, J. A., Knox, J. R., Zhao, H., Frère, J.-M. and Ghuysen, J.-M. (1989) J. Mol. Biol. **209**, 281–295

43 Mayo, M. A., Robinson, D. J., Jolly, C. A. and Hyman, L. (1989) J. Gen. Virol. **70**, 1037–1051

44 Amerik, A. Y., Antonov, V. K., Gorbalenya, A. E., Kotova, S. A., Rotanova, T. V. and Shimbarevich, E. V. (1991) FEBS Lett. **287**, 211–214

45 Rawlings, N. D., Pearl, L. H. and Buttle, D. J. (1992) Biol. Chem. Hoppe-Seyler **373**, in the press

46 Ohno, S., Emori, Y., Imajoh, S., Kawasaki, H., Kisaragi, M. and Suzuki, K. (1984) Nature (London) **312**, 566–570

47 Delaney, S. J., Hayward, D. C., Barleben, F., Fischbach, K. F. and Miklos, G. L. G. (1991) Proc. Natl. Acad. Sci. U.S.A. **88**, 7214–7218

48 Bazan, J. F. and Fletterick, R. J. (1988) Proc. Natl. Acad. Sci. U.S.A. **85**, 7872–7876

49 Gorbalenya, A. E., Donchenko, A. P., Blinov, V. M. and Koonin, E. V. (1989) FEBS Lett. **243**, 103–114

50 Hämmerle, T., Hellen, C. U. T. and Wimmer, E. (1991) J. Biol. Chem. **266**, 5412–5416

51 Yu, S. F. and Lloyd, R. E. (1991) Virology **182**, 615–625

52 Tang, J., James, M. N. G., Hsu, I. N., Jenkins, J. A. and Blundell, T. L. (1978) Nature (London) **271**, 618–621

53 Miller, M., Jaskólski, M., Rao, J. K. M., Leis, J. and Wlodawer, A. (1989) Nature (London) **337**, 576–579

54 Seelmeier, S., Schmidt, H., Turk, V. and Von Der Helm, K. (1988) Proc. Natl. Acad. Sci. U.S.A. **85**, 6612–6616

55 Tsuru, D., Shimada, S., Maruta, S., Yoshimoto, T., Oda, K., Murao, S., Miyata, T. and Iwanaga, S. (1986) J. Biochem. (Tokyo) **99**, 1537–1539

56 Takahashi, K., Inoue, H., Sakai, K., Kohama, T., Kitahara, S., Takishima, K., Tanji, M., Athauda, S. B. P., Takahashi, T., Akanuma, H., Mamiya, G. and Yamasaki, M. (1991) J. Biol. Chem. **266**, 19480–19483

57 Medina, J. F., Wetterholm, A., Rådmark, O., Shapiro, R., Haeggström, J. Z., Vallee, B. L. and Samuelsson, B. (1991) Proc. Natl. Acad. Sci. U.S.A. **88**, 7620–7624

58 Matthews, B. W., Weaver, L. H. and Kester, W. R. (1974) J. Biol. Chem. **249**, 8030–8044

59 Le Moual, H., Devault, A., Roques, B. P., Crine, P. and Boileau, G. (1991) J. Biol. Chem. **266**, 15670–15674

60 Bode, W., Gomis-Rüth, F. X., Huber, R., Zwilling, R. and Stöcker, W. (1992) Nature (London) **358**, 164–167

61 Vallee, B. L. and Auld, D. S. (1990) Biochemistry **29**, 5647–5659

62 Becker, A. B. and Roth, R. A. (1992) Proc. Natl. Acad. Sci. U.S.A. **89**, 3835–3839

63 Meulenberg, J. J. M., Sellink, E., Loenen, W. A. M., Riegman, N. H., van Kleff, M. and Postma, P. W. (1990) FEMS Microbiol. Lett. **71**, 337–344

64 DiBenedetto, A. J., Harada, H. A. and Wolfner, M. F. (1990) Dev. Biol. **139**, 134–148

65 Zwickl, P., Grziwa, A., Pühler, G., Dahlmann, B., Lottspeich, F. and Baumeister, W. (1992) Biochemistry **31**, 964–972

66 Barrett, A. J. (1986) in Proteinase inhibitors (Barrett, A. J. and Salvesen, G., eds.), pp. 3–22, Elsevier Science Publishers, Amsterdam

Received 29 July 1992; accepted 27 August 1992