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

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

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)

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)

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.