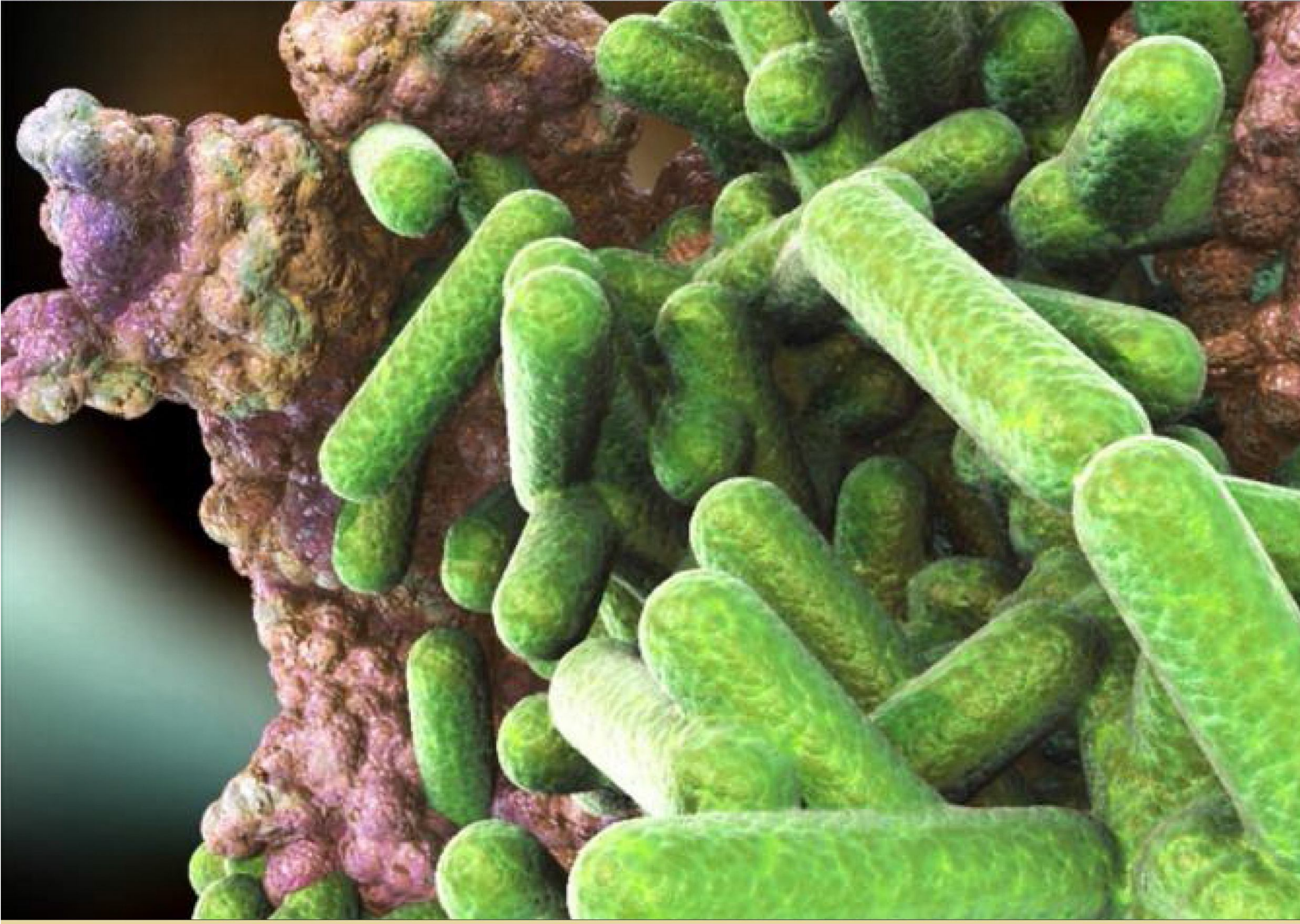


Rui Gonçalo Batista Mamede da Cruz

NEW RETROVIRAL-LIKE MEMBRANE-ASSOCIATED ASPARTIC PROTEASES FROM RICKETTSIAE: BIOCHEMICAL CHARACTERIZATION AND SPECIFICITY PROFILING

Tese de Doutoramento em Bioquímica, especialidade em Tecnologia Bioquímica, orientada pela Doutora Isaura Simões e pelo Professor Doutor Carlos Faro e apresentada ao Departamento de Ciências da Vida da Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Julho de 2014

· U C ·

Universidade de Coimbra

Novas proteases aspárticas membranares do tipo retroviral de rickettsiae: caracterização bioquímica e determinação de especificidade

New retroviral-like membrane-associated aspartic proteases from rickettsiae: biochemical characterization and specificity profiling

Tese de Doutoramento apresentada à Universidade de Coimbra para obtenção do grau de Doutor em Bioquímica (especialidade de Tecnologia Bioquímica).

Doctoral thesis in the scientific area of Biochemistry (specialty Biochemistry Technology) presented to the University of Coimbra.

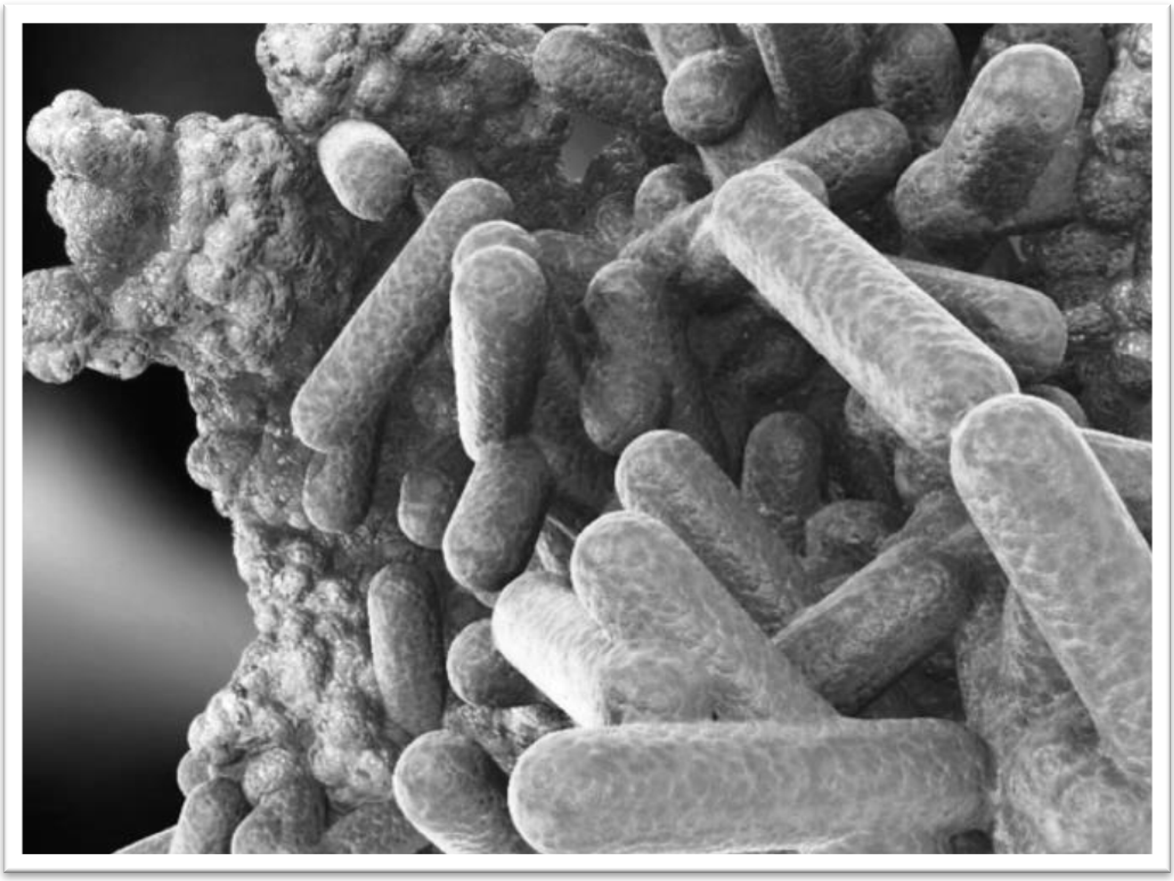
Rui Gonçalo Batista Mamede da Cruz

Coimbra - 2014

Department of Life Sciences
University of Coimbra

· U C ·

UNIVERSIDADE DE COIMBRA



Front cover image illustrates *Rickettsia* cells (in green) invading mammalian host cells.

Publishing rights were granted by Michael Taylor, the original author of the image.

Agradecimentos / Acknowledgments

A conclusão desta etapa não teria sido possível sem o apoio incondicional da família, amigos, orientadores, colegas e de todos aqueles que foram cruzando o meu caminho e deixando o seu contributo, em particular nos últimos quatro anos que conduziram a esta dissertação. Gostaria por isso de lhes deixar algumas palavras de profundo e sincero agradecimento.

Em primeiro lugar, gostaria de agradecer à Doutora Isaura Simões e ao Professor Carlos Faro, por me terem acolhido e proporcionado as excepcionais condições para a realização deste trabalho. Em especial à Doutora Isaura Simões, expresso a minha profunda gratidão pela orientação, rigor, sentido crítico e paixão com que encara a ciência. Agradeço-lhe por acreditar em mim e por me fazer ir sempre mais além.

Um agradecimento muito especial a todos aqueles que fazem e fizeram parte do laboratório de Biotecnologia de Molecular do Biocant, pelo seu contributo no meu crescimento científico, profissional e pessoal. Um agradecimento ao Pedro Castanheira pelo conhecimento que me transmitiu e pelas palavras de incentivo e encorajamento. À Carla Almeida um enorme obrigado pela disponibilidade e partilha de conhecimento e, acima de tudo, pelos sempre bons conselhos. Aos meus colegas de bancada Ana Sofia Lourenço, André Soares, Marisa Simões, Joana Furtado, Pedro Curto, Ana Rita Leal, Liliana Antunes e a todos os outros, agradeço pelos bons momentos de amizade, pela boa disposição, pelas estimulantes conversas e desabafos mútuos. Muito obrigado por tornarem este percurso mais agradável.

Ao Laboratório de Proteómica, de uma forma especial ao Bruno Manadas, à Vera Mendes e à Sandra Anjos, agradeço pela disponibilidade e colaboração nas análises de espectrometria de massa.

Aos restantes colegas e amigos que integram a família Biocant, o meu muito obrigado pelos bons momentos de convívio que em muito contribuíram para que que cada dia fosse encarado com particular motivação.

Uma palavra de carinho a todos os colegas e amigos do CNC (dos tempos do IBILI), pelas boas recordações dos bons momentos aí vividos e pelas amizades que ficaram para a vida. Um muito obrigado.

To Doctor Juan Martinez, who kindly welcomed me in his lab, and to Sean Riley, I deeply thank the knowledge and advice they shared with me throughout my stay, as well as the great opportunity to know a different laboratorial reality, and the meaningful personal experience it all meant. I extend my thanks to Abby and all members of Doctor Macaluso's Lab for making my life in Baton Rouge so enjoyable.

I would like also to thank Doctor Christopher Overall and his lab team, especially to Pitter Huesgen, for the guidance and all the support during my stay in Vancouver, for all the collaboration, shared knowledge and friendship that made that time in Vancouver an exceptional experience.

A todos os meus amigos, em particular ao Rui, um enorme obrigado por estarem sempre presentes e por me permitirem aproveitar o que de melhor a vida tem para oferecer. Muito obrigado pela amizade, companhia e afecto.

Agradeço à minha família todo o apoio incondicional. De uma forma muito especial, agradeço aos meus pais por me terem permitido ser o que sou hoje, pois tudo o que consegui a eles o devo. Agradeço-lhes por estarem ao meu lado em todos os momentos, pela preocupação constante e por serem um exemplo de integridade, humildade e trabalho. Ao meu irmão por ser um amigo e por tudo o que me ensinou ao longo da vida. Tenho um grande orgulho em ser vosso filho e irmão.

Por fim, um desmedido obrigado à Dulce por ser o meu porto seguro. As palavras serão sempre poucas para expressar tudo aquilo que representas para mim. Agradeço-te acima de tudo por seres parte daquilo que eu sou. As minhas conquistas serão sempre igualmente tuas...

The work herein presented is part of the manuscript entitled, *RC1339/APRc from Rickettsia conorii is a Novel Aspartic Protease with Properties of Retropepsin-Like Enzymes,* accepted for publication at PLOS Pathogens (manuscript reference: PPATHOGENS-D-14-00592).

Rui Cruz, Pitter Huesgen, Sean P. Riley, Alexander Wlodawer, Carlos Faro, Christopher M. Overall, Juan J. Martinez and Isaura Simões. 2014*. RC1339/APRc from Rickettsia conorii is a Novel Aspartic Protease with Properties of Retropepsin-Like Enzymes*. PLOS Pathogens (DOI: 10.1371/journal.ppat.1004324).

Este trabalho foi financiado por Fundos FEDER através do Programa Operacional Factores de Competitividade – COMPETE e por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto PTDC/SAU-MII/107942/2008; *Rui Cruz usufruiu de uma bolsa de Doutoramento com a referência SFRH/BD/47638/2008.*

Table of contents

Resumo

Resumo

Os membros do género *Rickettsia* são bactérias intracelulares obrigatórias do tipo gram-negativas, cuja transmissão a mamíferos pode ocorrer através de vetores artrópodes como carraças, pulgas ou piolhos. Entre as várias espécies identificadas, muitas são patogénicas para o Homem causando doenças infeciosas agudas das quais se destacam o tifo epidémico (*Rickettsia prowazekii*), a febre maculosa das montanhas rochosas (*Rickettsia rickettsii*) e a febre escaro-nodular (*Rickettsia conorii*). A elevada patogenicidade e o caráter emergente destas doenças, associados à inexistência de vacinas eficazes para a sua prevenção, reforçam inequivocamente a necessidade de identificar novos fatores proteicos para o desenvolvimento de terapêuticas inovadoras.

Neste sentido, tem-se assistido nas últimas décadas a avanços significativos na compreensão dos mecanismos de patogénese e de resposta imunitária às rickettsioses. Contudo, a validação da função biológica de genes de *Rickettsia* tem sido amplamente limitada pela natureza estritamente intracelular destes organismos que dificulta a sua manipulação. Por conseguinte, a comparação entre os múltiplos genomas disponíveis de *Rickettsia* tem revelado ser o método mais expedito para a identificação de novos fatores proteicos potencialmente implicados na patogenicidade destes micro-organismos.

Este trabalho descreve a identificação e caracterização de uma nova protease membranar do tipo retropepsina, altamente conservada em 55 genomas de *Rickettsia*. Apesar da baixa similaridade na sequência de aminoácidos relativamente a outras retropepsinas, demonstrámos que a proteína codificada pelo gene homólogo RC1339 de *R. conorii* Malish 7, designada por APRc para protease aspártica de *Rickettsia conorii*, é uma enzima ativa com propriedades altamente reminiscentes desta família de proteases aspárticas. Entre outras, destacam-se a atividade autolítica comprometida pela mutação do aspartato catalítico, a acumulação na forma dimérica, uma atividade ótima a pH de 6 e a inibição por inibidores específicos da protease do vírus da imunodeficiência humana do tipo 1. Além disso, utilizando uma abordagem de mapeamento de especificidade de alto débito, foi possível confirmar que os determinantes de especificidade da APRc são semelhantes aos de outras proteases aspárticas de ambos os tipos, retropepsina e pepsina.

Neste trabalho, demonstrámos também que o gene codificante da APRc é transcrito e traduzido em pelo menos duas espécies patogénicas de *Rickettsia* (*R. conorii* e *R. rickettsii*), e que esta proteína é integrada na membrana externa de ambas. Ao explorar as potenciais funções biológicas da APRc, verificámos que esta protease catalisa o processamento *in vitro* de

dois membros da família das proteínas autotransportadoras envolvidas na adesão e invasão de *Rickettsia*: Sca5/rOmpB e Sca0/rOmpA. Estes resultados apontam assim para a participação da APRc numa via proteolítica relevante para a virulência destes micro-organismos, surgindo esta protease como um alvo interessante para a intervenção terapêutica contra as rickettsioses.

Por fim, ao demonstrar que a APRc é um novo membro da família das proteases aspárticas do tipo retropepsina, provamos simultaneamente que estas enzimas estão efetivamente presentes em bactérias gram-negativas intracelulares, pelo que poderão representar uma forma ancestral desta classe de proteases.

Abstract

Abstract

Members of the genus *Rickettsia* are obligate intracellular, gram-negative, arthropod-borne pathogens of humans and other mammals, causing severe infections including epidemic typhus (*Rickettsia prowazekii*), Rocky Mountain spotted fever (*Rickettsia rickettsii*), and Mediterranean spotted fever (*Rickettsia conorii*). The life-threatening character of diseases caused by many *Rickettsia* spp. and the lack of reliable protective vaccine against rickettsioses strengthens the importance of identifying new protein factors for the potential development of innovative therapeutic tools. However, progress in correlating rickettsial genes and gene functions has been greatly hampered by the intrinsic difficulty in working with these obligate intracellular bacteria, despite the increasing insights into the mechanisms of pathogenesis of and the immune response to rickettsioses. Therefore, comparison of the multiple available genomes of *Rickettsia* is proving to be the most practical method to identify new factors that may play a role in pathogenicity.

The present work reports the identification and characterization of a novel membrane-embedded retropepsin-like homologue, highly conserved in 55 *Rickettsia* genomes. Using *R. conorii* Malish 7 gene homologue *RC1339* as our working model we demonstrate that, despite the low overall sequence similarity to retropepsins, the gene product of *RC1339* APRc (for Aspartic Protease from *Rickettsia conorii*) is an active enzyme with features highly reminiscent of this family of aspartic proteases, such as autolytic activity impaired by mutation of the catalytic aspartate, accumulation in the dimeric form, optimal activity at pH 6, and inhibition by specific HIV-1 protease inhibitors. Moreover, specificity preferences determined by a high-throughput profiling approach confirmed common preferences between this novel rickettsial enzyme and other aspartic proteases, both retropepsin and pepsin-like enzymes. Additionally, we have also shown that APRc is transcribed and translated by at least two pathogenic rickettsial species, *R. conorii* and *R. rickettsii*, and is integrated into the outer membrane of both species.

By further exploring one of its putative biological roles, we have demonstrated that APRc is sufficient to catalyze the *in vitro* processing of two conserved high molecular weight autotransporter adhesin/invasion proteins, Sca5/rOmpB and Sca0/rOmpA, thereby suggesting the participation of this enzyme in a relevant proteolytic pathway in rickettsial virulence. As a novel *bona fide* member of the retropepsin family of aspartic proteases, APRc emerges as an intriguing target for therapeutic intervention against fatal rickettsioses.

Finally, with this work we demonstrate that retropepsin-type aspartic proteases are indeed present in gram-negative intracellular bacteria such as *Rickettsia*, suggesting that these enzymes may represent an ancestral form of this class of proteases.

Abbreviations

Abbreviations

ACE: Angiotensin-Converting Enzyme

AG: Ancestral Group

AP: Aspartic Protease

BACE1: β-secretase 1

BCA: Bicinchoninic Acid Assay

BLV: Bovine Leukemia Virus

CA: Capsid

CDR1: Constitutive Disease Resistance 1

Ddi1: DNA-damage Inducible Protein 1

DSS: Disuccinimidyl Suberate

EIAV: Equine Infectious Anemia Virus

FIV: Feline Immunodeficiency Virus

FV: Foamy Virus

GPR: Germination Protease

GST: Glutathione S-Transferase

HFV: Human Foamy Virus

HIV-1: Human Immunodeficiency Virus-1

HIV-2: Human Immunodeficiency Virus-2

HTLV-1: Human T-lymphotropic virus 1

IN: Integrase

LC-MS/MS: Liquid Chromatography–Tandem Mass Spectrometry

MA: Matrix

MAV: Myeloblastosis Associated Virus

MMLV: Moloney Murine Leukemia Virus

MMTV: Mouse Mammary Tumor Virus

MPMV: Mason- Pfizer Monkey Virus

MSF: Mediterranean Spotted Fever

NC: Nucleocapsid

PCS1: Promotion of Cell Survival 1

PFA: Paraformaldehyde

PICS: Proteomic Identification of protease Cleavage Sites

PLA2: Phospholipase A2

PLD: Phospholipase D

PR: protease

qPCR: Real-time Quantitative Polymerase Chain Reaction

RH: RNaseH

RMSF: Rocky Mountain Spotted Fever

RP-HPLC: Reversed-Phase High Performance Liquid Chromatography

RT: Reverse Transcriptase

SAM: Sorting and Assembly Machinery

SASPase: Skin Aspartic Protease

Sca: Surface cell antigen

SEC: Size-exclusion Chromatography

SFG: Spotted Fever Group

SFVmac: Simian Foamy Virus from macaques

SPATEs: Serine Protease Autotransporters

TFA: Trifluoroacetic Acid

TFPP: Type IV Prepilins Peptidase

TFR: Transframe Region

TG: Typhus Group

THP1: human monocytic leukemia cell line

TlyC: Hemolysin C

TM: Transmembrane

TMH: Transmembrane α-helix

TOM: Translocase of the Outer Membrane

WDSV: Walleye Dermal Sarcoma Virus

XMRV: Xenotropic Murine Leukemia Virus-related Virus

Chapter I. Introduction

Chapter I. Introduction

1.1. Proteolytic enzymes

General description

Proteins are the most complex and functionally diverse macromolecules in living organisms. Among the many different types of proteins, enzymes have specific catalytic properties responsible for accelerating chemical reactions within cells and, thus, are essential for sustaining life. Depending on the type of catalytic activity they have, enzymes are divided into different classes. One of such classes comprises proteases, also termed peptidases, which are characterized by their capacity to selectively catalyze the hydrolysis of amide bonds within peptides and proteins from monomers to multimeric complexes.

Encoded by approximately 2% of the genes in all kinds of organisms, proteases constitute one of the largest functional groups of proteins[1]. Regardless of the complexity of the organism, proteases control a wide range of biological functions and processes, including protein turnover, cell growth, cell death, immune defense and secretion[2–6]. Consistent with these essential roles, many proteases are involved in human diseases, ranging from degenerative and inflammatory diseases to cancer[7–9]. In plants, proteases also play key roles in a striking diversity of biological processes including embryogenesis, gametophyte survival, chloroplast biogenesis, stomata development, and local and systemic defense responses[10]. Likewise, many pathogenic viruses and bacteria use proteases for their life cycle or as virulence factors for infection of host cells[6,11]. Finally, proteases also occupy a pivotal position with respect to their numerous practical applications in the biotechnological industry as biochemical reagents or in the manufacture of numerous products[12–14].

Proteases have evolved as important regulatory enzymes which, depending on their specificity, can modify proteins post-translationally at highly specific sites (limited proteolysis) for activation and maturation of proteins or removal of signal or transit peptides. On the other hand, housekeeping proteases are responsible for unspecific and total degradation of damaged, misfolded and potentially harmful proteins, providing free amino acids for the synthesis of new proteins. Proteases can be divided into exoproteases, whose activity is directed at the amino or carboxyl termini of polypeptide chains, or endoproteases, which preferentially cleave peptide bonds in the inner regions of proteins.

The *MEROPS* database is the most modern and organized system that provides an evolutionary hierarchical classification of proteases into classes, families and clans[1]. Accordingly, based on structural and catalytic homology, proteases are categorized into the following nine classes: serine, aspartic, cysteine, metalloproteases, glutamic, threonine, asparagine, mixed catalytic type, and the ninth class, which comprises a number of proteases that cannot yet be assigned to any particular catalytic type. Each class of proteases is specific in its ability to break a certain peptide bond and displays a characteristic set of functional amino acid residues arranged in a specific configuration to produce its catalytic site[1]. To define the position of the substrate residues interacting with the protease substrate-binding subsites, a general nomenclature was formulated by Schechter and Berger[15]. According to this nomenclature, the protease-binding subsites (S) for residues located at the N-terminal side (prime side) of the scissile peptide bond are designated as S1, S2 and so on, whereas the corresponding substrate peptide (P) residues are designated as P1, P2 and so on. Binding subsites and substrate residues from the C-terminal side (non-prime side) of the scissile peptide bond are designated as S1′, S2′ and P1′, P2′ and so on, respectively.

Aspartic proteases

Aspartic proteases (APs) are a ubiquitous class of enzymes, which use a pair of highly conserved aspartate residues in the active site to activate a water molecule and hydrolyze peptide bonds. According to the *MEROPS* database, APs are currently classified into 16 families, which are in turn included into 6 clans[1], as depicted in Table 1. The families differ according to the conserved residues for the enzymatic functionality, the position of the catalytic aspartic acid residues in the peptide chains, substrate specificity, the number of disulfide bridges in their structure and the optimal pH at which the enzymes function[1]. Despite these variations, conserved sequence motifs have been identified for the majority of APs. For instance, the catalytic aspartate residues of all members of clan AA are organized in the consensus motif Asp-Thr/Ser-Gly, contained in the sequence Xaa-Xaa-Asp-Xbb-Gly-Xcc, where Xaa is hydrophobic, Xbb is Thr or Ser, and Xcc is Ser, Thr or Ala.

Table 1. *Clans and families of aspartic proteases.*

CLAN	FAMILY	TYPE PROTEASE
AA	A1	Pepsin A (*Homo sapiens*)
	A2	HIV-1 retropepsin (human immunodeficiency virus 1)
	A3	Cauliflower mosaic virus-type peptidase (cauliflower mosaic virus)
	A9	Spumapepsin (human spumaretrovirus)
	A11	Copia transposon peptidase (*Drosophila melanogaster*)
	A28	DNA-damage inducible protein 1 (*Saccharomyces cerevisiae*)
	A32	PerP peptidase (*Caulobacter crescentus*)
	A33	Skin SASPase (*Mus musculus*)
AC	A8	Signal peptidase II (*Escherichia coli*)
AD	A22	Presenilin 1 (*Homo sapiens*)
	A24	Type 4 prepilin peptidase 1 (*Pseudomonas aeruginosa*)
AE	A25	GPR peptidase (*Bacillus megaterium*)
	A31	HybD peptidase (*Escherichia coli*)
AF	A26	Omptin (*Escherichia coli*)
UNASSIGNED	A5	Thermopsin (*Sulfolobus acidocaldarius*)
	A36	Sporulation factor SpoIIGA (*Bacillus subtilis*)

The largest family of APs is the pepsin family (A1, clan AA) with 5277 identified sequences, which is further divided into subfamily A1A of pepsin-like enzymes (also called pepsins, hereafter) and subfamily A1B of plant APs; the second biggest family is the retropepsin family (A2, clan AA) with 719 identified sequences. The members of families A1 and A2 are known to be related to each other, with those of family A3 also showing some correlation to A1 and A2 members. With exception for families A5 and A36 which have not yet been assigned to any clan, the remaining families (A28; A22; A24; A25; A31; A26) were included into clans AC, AD, AE and AF, which greatly differ regarding the catalytic motif composition and overall structural organization, with the majority being partially or totally membrane-embedded proteases.

The far most characterized APs belong to families A1 and A2 (pepsin and retropepsin-like enzymes, respectively) as the result of the enormous interest they have received, given their important roles in some human diseases. The present chapter focuses specifically on these two families of APs, with a detailed description and comparison of their general features, in order to provide a context in which to compare the specific examples found in the following chapters.

Distribution and relevance

APs have been widely described as having important functional roles in a multitude of organisms, from vertebrates to fungi, plants and retroviruses, and more recently homologues have been reported from prokaryotes (Table 2)[16,17]. Amongst all members of the pepsin family found in mammalian, the most studied are involved in digestion and protein degradation, such as pepsin - which is one of the principal proteolytic enzymes in the digestive system[18] - and chymosin[19], which has been used for thousands of years in cheese making, as well as the major lysosomal enzyme cathepsin D[20]. Nevertheless, other mammalian APs such as renin[21], which plays an important role in blood pressure, or β-secretase 1 (BACE1/Memapsin 2)[7,22] implicated in Alzheimer's disease, have also been the subject of intensive study over the last decades. Secreted aspartic peptidases (or SAP) of *Candida* spp., and plasmepsins I and II, found in the malarial parasite *Plasmodium falciparum*, are two other groups of pepsin-like APs that have also been investigated in detail as key targets for fungal and parasitic infections[23,24].

In plants, pepsin-like APs have been found in seeds, suggesting a role in the processing of storage proteins during ripening and germination; in leaves, indicating a role in defense mechanisms against pathogens, and in flowers where they are suggested to play a role in sexual reproduction[25]. In addition, some plant APs have also been implicated in cell death events and response to stress (e.g., CDR1, PCS1, UNDEAD)[26].

Proteases of retroviruses, such as leukemia viruses, immunodeficiency viruses (e.g., HIV-1), infectious anemia viruses, and mammary tumor viruses, form an important family (A2) with those encoded by several endogenous viral sequences in primates and retrotransposons in yeast and *Drosophila*[27]. These retroviral proteases (PRs) (hereafter also termed as retropepsins), represented by the far most characterized HIV-1 PR, are critical enzymes in viral propagation. They are initially synthesized with other viral proteins as polyprotein precursors (Gag and Gag-Pol) which are subsequently cleaved by the viral protease activity at precise sites to produce the functional proteins and enzymes[28]. Additionally, it has also been shown that many host proteins are also substrates of HIV-1 PR which can contribute to the pathogenicity of the virus[11].

Table 2. *Examples of aspartic proteases and their biological functions.*

ASPARTIC PROTEASES	FAMILY	BIOLOGICAL FUNCTIONS
Human aspartic proteases		
Pepsin	A1	Protein digestion in the stomach
Gastricsin	A1	Digestion in stomach and seminal plasma
Cathepsin D	A1	General protein degradation and turnover
Napsin	A1	Suppression of cancer growth
Renin	A1	Regulation of blood pressure
Memapsin 1 (BACE2)	A1	Insulin receptor trafficking and signaling
Memapsin 2 (BACE1)	A1	Neuronal development (formation of myelin sheaths)
Presenilin	A22	Cellular differentiation and proteolysis of membrane proteins
Type IV prepilin peptidase	A24	Type IV pilus formation and protein secretion
Plant aspartic proteases		
Cardosins	A1	Plant sexual reproduction and postembryonic development
Phytepsin	A1	Protein storage processing
CDR1	A1	Disease resistance signaling
PCS1	A1	Prevention of cell death
UNDEAD	A1	Regulation of programmed cell death
Microbial and pathogen aspartic proteases		
Shewasin A	A1	Not determined
Plasmepsin 2	A1	Hemoglobin digestion in vacuoles of *Plasmodium falciparum*
Rhizopuspepsin	A1	Extracellular protein hydrolysis by *Rosa chinensis*
Endothiapepsin	A1	Extracellular protein hydrolysis by *Eutypella parasitica*
Penicillopepsin	A1	Extracellular protein hydrolysis by penicillium fungi
Candida SAPs	A1	Virulent factors for *Candida* species
Barrierpepsin	A1	Cleavage of α-factor for cell cycle regulation of yeast
Retroviral aspartic proteases		
HIV-1 PR	A2	Processing of viral polyproteins for virion assembly
HTLV-1 PR	A2	Processing of viral polyproteins for virion assembly

Structure and processing activity

Up to date, tertiary structures solved for APs show a unique protein fold unrelated to that of any other protease. The first AP to be sequenced was porcine pepsin, by Tang *et al.* in 1973[29]. Ten years later, James and Sielecki published the first 3D structure of an AP (penicillopepsin)[30], while the second structure to be analyzed by X-ray diffraction was from pepsin[31]. Proteases of retroviruses such as Rous sarcoma virus (RSV) and HIV-1 have also been extensively studied and their crystal structures have been determined as early as 1989[32–35].

From these primordial studies, it became clear that all mature pepsin-like enzymes have a considerable degree of structural similarity. They are bilobal molecules, containing two topological similar N and C domains predominantly formed by β-sheets and related by a pseudo 2-fold axis, with the active-site cleft located between the lobes with about 35-40 Å long (Figure 1A). These two homologous domains are linked by a six-stranded, antiparallel β-sheet that is the only structured motif shared by the two lobes. Each lobe contributes with one catalytic aspartic residue (Asp32 and Asp215, pepsin numbering) located within the hallmark motif Asp-Thr/Ser-Gly[36]. This motif is followed by a hydrophobic-hydrophobic-Gly conserved sequence which, together with the catalytic sequence motif, forms a structural feature known as psi-loop. These psi-loop/alpha-helix motifs fix the central structure of the enzyme and thereby define the catalytic machinery of APs[36,37].

Although the two lobes of pepsins are structurally similar, an extended β-hairpin loop on the N-terminal lobe surface covers the binding cleft at the junction of the two lobes to form a hinged flexible flap region that encloses substrates or inhibitors into the active site (switching between an open or closed conformation). The inspection of different available 3D structures of pepsin-like enzymes in the free form or complexed with inhibitors has shown that the conserved residues (Tyr75, Gly76 and Thr77 in pepsin sequence) located at the tip of the flap influence ligand selectivity by interaction with amino acids in the P1-P2' positions[38,39].

Other landmark residues and motifs are characteristic of the members of pepsin family, including several intramolecular disulfide bonds at characteristic locations, whose number may vary from sequence to sequence and may impact on the stability of the native-state of each protein[36,37,40].

Due to the conservation of the catalytic motif Asp-Thr/Ser-Gly, structural similarities and other biochemical features, retroviral proteases have been promptly predicted to be related with members of pepsin family[36,41]. Nevertheless, the molecular architecture of retroviral proteases is distinct among enzymes, with no other known examples of active sites formed in a similar manner[41]. In contrast to pepsins, retroviral proteases are β-homodimers consisting of two chemically identical subunits in a nearly 2-fold internal symmetric arrangement, each one with 99-138 residues and a molecular weight ranging between 11-15 kDa (Figure 1B). A remarkable finding was the observation that despite the low sequence similarity, each of these subunits of retropepsins is structurally similar to a single domain of the pepsin-like enzymes[42]. Moreover, the secondary structure of all retroviral proteases follows a structural template where the monomeric molecule is formed by duplication of four structural elements. Figure 1C depicts the secondary structure of RSV PR, the first crystal structure of a retroviral protease to be determined[35], consisting of a hairpin, a wide loop containing the catalytic Asp residue, an α-helix and lastly a second hairpin[35,41]. The first HIV-1 PR structure was determined afterwards at Merck laboratories with a 3 Å resolution[32]. The active site of retroviral proteases is formed within the interface of the two monomers, where each monomer contributes with one conserved aspartic acid for active site assembling (Asp25 and Asp25', HIV-1 PR numbering).

Accordingly, retropepsins are only active in the form of noncovalently bound homodimers. Resembling pepsins, these enzymes also retain the hydrophobic-hydrophobic-Gly sequence, maintaining the conserved psi-loop in each monomer.

Structural studies have shown that the N- and C-terminal regions of the two monomers of HIV-1 PR form a structure of four layered β-strands, which constitutes the interface between both subunits and contribute significantly to dimerization[43,44]. As a prerequisite for dimerization to occur it was also observed that, under physiological conditions, monomeric HIV-1 PR subunits are partially structured exhibiting a similar secondary and tertiary structure of a single subunit of the active dimer[43,44]. Moreover, instead of the single flap observed in pepsin-like enzymes, the homodimeric retroviral proteases have two much less structured flaps (than that found in pepsins) that interact not only with the substrate but also with each other[39,45,46]. Nuclear magnetic resonance studies and comparison of crystal structures of HIV-1 PR in a free form or bound to substrates or inhibitors have shown a complex dynamic behavior of flaps which alternate between three conformations. Free protease exists primarily in the "semi-open" conformation but transiently changes to the fully "open" conformation allowing the ligand to enter into the active site; ligand induces the closing of the flaps into a "closed" conformation and it is converted again to the "semi-open" conformation upon removal of ligand[45–47]. Retroviral proteases flap tips usually contain several glycines (instead of the single glycine in pepsin-like APs), identified to be important in the flap closing mechanism which is conserved across known structures of the retropepsin family[45]. Importantly, the prominent differences between the flap functional regions from pepsins and retropepsins, associated with their different mechanisms of molecular recognition and binding between the enzyme and the substrate, have been recognized to have a major impact on the specificity of each AP[45,48].

An important feature of all APs is a quite rigid network of hydrogen bonds occurring at the active site, called the "fireman's grip". This structure involves the hydroxyl groups of each Thr/Ser residues in the active site Asp-Thr/Ser-Gly triplets (Thr26 in the case of HIV-1 PR), which accepts a hydrogen bond from the main-chain amine group of the Thr26 in the opposing loop and also donates a hydrogen bond to the oxygen atom of the carbonyl group of residue 24 on the opposite loop[27]. The aspartic acid residues are bridged by a water molecule, located within hydrogen-bonding distance of the oxygen atoms of the Asp25 carboxyl group[27,49,50]. Although provided that the interactions in the N- and C-terminal dimerization domain and the flap region are responsible for stabilizing the dimer, it was later found that this complex scaffold of hydrogen bonds also aids to dimerization by mediating the initial contact of the two monomer molecules and by adjusting them to the proper conformation and/or orientation[49,50].

Importantly, upon the binding of substrates with asymmetric shapes, the protease adapts and its symmetry is lost[51].

Although the active-site catalytic motif is common to all APs, the sequence Gly86-Arg87-Asn/Asp88 (HIV-1 PR numbering) in the α-helix is unique to retroviral proteases. The Arg87 residue plays a crucial role in the stability of the dimer, as the loss of hydrogen bond between this residue and Asp29 results in the destabilization of the dimer interfaces, particularly between the C-terminal β-strands[52].

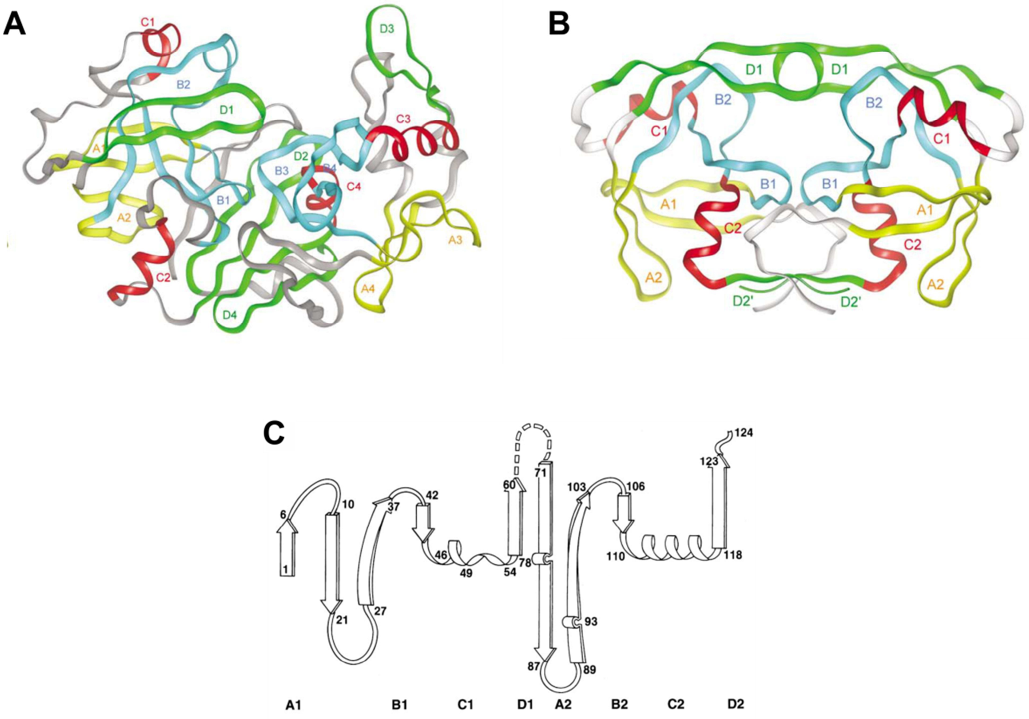


Figure 1. *Structural template for pepsin-like and retroviral proteases.* (**A**) In pepsin-like enzymes, the hairpin loops A1 and A2 are labeled in the left-side domain and A3 and A4 in the right-side domain. (**B**) In the symmetrical retroviral dimer, the corresponding loops A1 and A2 are shown in yellow in each monomer, shown as stereo pairs. Likewise, the psi-loop identified with B1 and loop B2 are shown in blue in each monomer of the retroviral dimer in (**B**), and the analogous loops in blue are labeled B1, B2, B3, and B4 in the single-chain enzymes. Loops B1 in the retroviral protease and B1 and B3 in the single-chain enzymes contain the catalytic residues. Helical segments C1 and C2 (red) in (**A**) are mirrored by segments C1-C4 in (**B**). Finally, loop D1 in the retroviral protease monomers provides a double flap structure in (**B**), whereas the 'half loops' D2 provide the four strands that form a β-sheet at the bottom of the dimer. In (**A**), loop D1 provides the flap on one side only, whereas D3 on the other side is pointing outward. Loops D2 and D4 provide the center of the β-sheet at the bottom of these enzymes. (**C**) Diagram of the secondary structure of retroviral proteases, with residue numbers corresponding to the structural elements observed in RSV PR. Adapted from Dunn *et al.*[27] and Wlodawer and Gustchina[41].

Despite this overall structural organization of mature APs from A1 and A2 families, pepsin-like enzymes - and it is believed most of the other A2 family members - are initially synthesized as inactive zymogens. Precursor forms of pepsins consist of the intact protease with an N-terminal extension comprising a signal peptide (pre-segment) and an activation segment (pro-segment) which assists in folding and prevents a premature activation of the enzyme (Figure 2A)[53]. While the highly hydrophobic pre-segment is removed upon entry of the zymogen in the endoplasmic reticulum, the post-translational removal of the pro-segment occurs typically by auto-proteolysis to liberate the active enzyme[54]. Physiological examples include the conversion of pepsinogens, prograstricsins and prochymosins into pepsins, grastricsins and chymosins, respectively[55]. For each type of zymogen, different isoforms (named isozymogens) have been identified and they are thought to be derived from different genes or from post-translational modifications such as phosphorylation and glycosylation[55]. Focusing on the most extensively studied example, the pepsinogen A, it comprises a pro-segment of 44 amino acids organized in one β-strand and three α-helices[36]. At neutral and alkaline pH conditions, this pro-segment binds and stabilizes the active site cleft with the aid of electrostatic, hydrophobic interactions and hydrogen bonding[36]. The self-cleavage of the pro-segment, i.e., the conversion of the zymogen into the respective active enzyme, occurs upon exposure to an acidic environment. Due to pH decrease, acidic residues get protonated resulting on the disruption of electrostatic interactions with positively charged amino acid residues of the pro-segment. Subsequent conformational rearrangements of both the active enzyme moiety and the pro-segment, lead to the proper positioning of each scissile bond in order to be cleaved by the exposed active site aspartates and, ultimately to the removal and dissociation of the pro-segment from the active center of the enzyme[56,57]. Although intramolecular proteolysis is the most common mechanism for AP activation, in some APs the activation occurs by intermolecular cleavage. For example, the activation of zymogens of APs involved in highly regulated physiological activities, such as proBACE1 and prorenin, is often accomplished by other proteases[22,36,58].

Similarly, retroviral proteases are initially synthesized as part of large polyproteins and must be processed during the maturation process of the virus in order to be active. Depending on the type of virus, the PR encoded in the so called Pro gene might be either produced in frame with the Gag (Gag-Pro) or Pol (Gag-Pro-Pol) polyproteins as in case of myeloblastosis associated virus (MAV)[59], with a stop codon suppression as exemplified by moloney murine leukemia virus (MMLV) (Gag-Pro)[60], by a translational frameshift mechanism as typified by HIV-1 (Gag-Pro-Pol)[61], or by a splicing event as it is the case of the human foamy virus (HFV) (Pro-Pol)[62]. Analyzing in closer detail the HIV-1, the translational frameshift within the p6 region allows translation beyond the p6 *gag* gene, resulting in a Gag-Pro-Pol fusion protein. Since the catalytic activity of retroviral proteases requires dimer formation, the first critical step in PR maturation involves the folding and dimerization of two PR domains in the form of Gag-Pro-Pol precursor in order to catalyze the hydrolysis of the peptide bonds at its termini[48]. The examination of the *in vitro* processing of a full-length Gag-Pro-Pol precursor confirmed that the initial cleavages carried out by the activated HIV-1 PR are intramolecular[63]. The released mature PR is critical for the virus assembly, maturation and propagation, as it is required for the proteolytic processing of the Gag and Gag-Pro-Pol precursors into functional structural proteins and enzymes. The Gag polyprotein is processed into its four mature protein domains: matrix (MA or p17), capsid (CA or p24), nucleocapsid (NC or p7) and p6 or transframe region (TFR or p6*), whereas the processing of the Pro-Pol segment creates the viral enzymes reverse transcriptase (RT), RNase H (RH), integrase (IN) and PR itself, in a process concomitant with particle release (Figure 2B).

A plausible pathway for the regulation of this viral PR emerged considering different *in vivo* and *in vitro* studies of precursor processing. *In vivo* studies showed that, depending on the pH, the autoprocessing can occur either stepwise in the order denoted in Figure 2B by sites 1, 2, and 3 at pH>5 or directly at site 3 at pH<5[64]. It has been also observed that prior to the cleavage at its N terminus (TFR/PR site), which precedes the cleavage of the C-terminal site (PR/RT), the dimer dissociation constant (Kd) of the protease is high and likely modulated by the TFR[65]. The high Kd is essential to allow effective recruitment of polyproteins at the plasma membrane, prior to the onset of polyprotein processing[65]. Interestingly, the activity of retroviral proteases appears to be in a good correlation with the way they are synthetized. In fact, while retroviral proteases with high specific activity are produced by frameshifting or stop codon suppression, thereby representing only 5-20% amount compared to the Gag polyprotein, the MAV PR is produced in frame of Gag and therefore, being in an equivalent amount with this substrate, has a substantially lower specific activity[66,67].

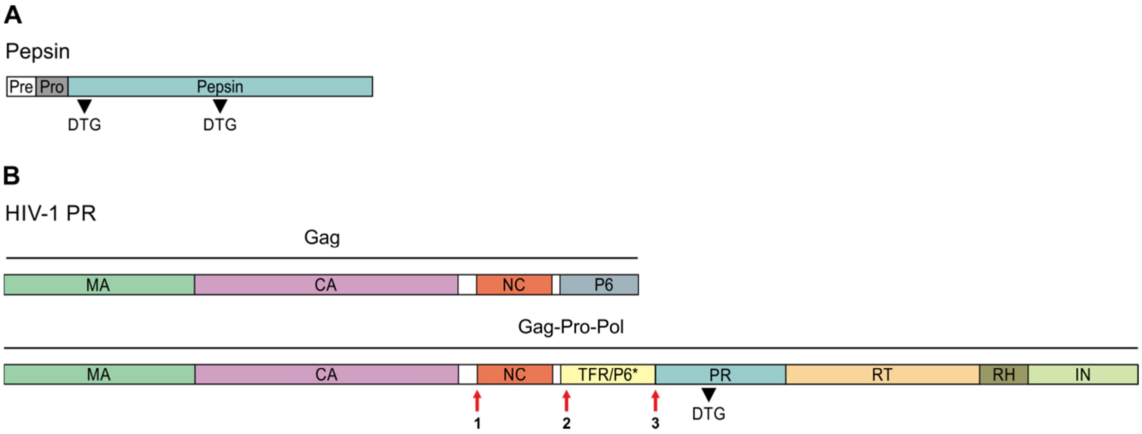


Figure 2. *Schematic representation of the domain organization of APs' precursors.* (**A**) The large majority of APs members of family A1 display a similar organization to pepsin with a signal peptide (Pre) and a pro-segment (Pro) at the N-terminus of the protease domain and the two catalytic aspartate residues (Asp32 and Asp215, pepsin numbering) contained in the typical Asp-Thr-Gly motifs. (**B**) For retroviral proteases such as HIV-1 PR, the chain containing the single Asp-Thr-Gly motif is comprised on the Gag-Pro-Pol polyprotein, which results from a translational frameshift within the p6 region. Gag-Pro-Pol polyprotein include the structural proteins p17 matrix (MA), p24 capsid (CA) and p7 nucleocapsid (NC) and the viral enzymes protease (PR), reverse transcriptase (RT), RNase H (RH) and integrase (IN). Adapted from Dunn *et al.*[27].

Catalytic mechanism, specificity and inhibition

Even though the overall fold of retropepsins and pepsins are different, they feature two conserved similar regions which are related to specific mechanistic functions (the catalytic pocket site and the residues in the β-sheets linking the two lobes) and thus, a common catalytic mechanism has been proposed for both families and extended for the remaining families from clan AA[27,36,37]. Mature APs are considered to follow a catalytic mechanism which begins with the binding of the substrate to form a loose complex followed by transition from the open to the closed conformation of flaps which set the substrate into the correct geometry for the catalytic process. After the bond cleavage event, the flaps open and the protease release the products with the concomitant re-forming of catalytic activity in the active site. Importantly, APs bind their substrates through hydrogen bond interactions with the substrate peptide backbone and by electrostatic and hydrophobic contacts between the side chains of substrate and well-defined pockets within the active site[68]. Catalysis of peptide bond hydrolysis by APs is dependent on the nucleophilic attack of an activated water molecule to the carbonyl carbon. A remarkable property of this catalytic center is its adaptation over a wide range of pH from pH 2 up to pH 7, although a maximal activity is generally observed at low pH (pH 3 to 4)[37].

Regarding the catalytic mechanism, it is established that the peptide bond cleavage catalyzed by APs follow a general acid–base mechanism, with the formation of a non-covalent neutral tetrahedral intermediate[36,69]. The microscopic details of this mechanism, on the contrary, are still under debate. Taking the different 3D structures of pepsins complexed with pepstatin A as models for the tetrahedral intermediate, different mechanisms have been proposed[69–71]. Among them, the most well accepted and representative one is the general acid-base mechanism proposed by Davies[72] (Figure 3), where the aspartate carboxyl groups act alternately as general acid and general base. According to this mechanism, the Asp on the C-lobe acts as a general base to remove one proton from the water molecule followed by nucleophilic attack of the water molecule to the carbonyl carbon of the substrate scissile bond, while the Asp on the N-lobe, which is at first in a nonionized form, donates its proton to the oxygen atom of the carbonyl of the substrate. An oxyanion tetrahedral intermediate is formed with the N-lobe Asp being hydrogen bonded to the attacking oxygen atom, while the hydrogen remaining on that oxygen is hydrogen bonded to the inner oxygen of C-lobe Asp. Next, a reversal of configuration occurs around the nitrogen atom of the scissile bond with the transfer of the hydrogen from the N-terminus Asp to that nitrogen atom. At the same time a proton is transferred from the inner oxygen of C-lobe Asp to the carbonyl oxygen on the peptide bond being cleaved. Hereafter the C-N bond breaks releasing the two products. The N-terminus Asp is negatively charged at this stage and ready for the next round of catalysis[36,70,71].

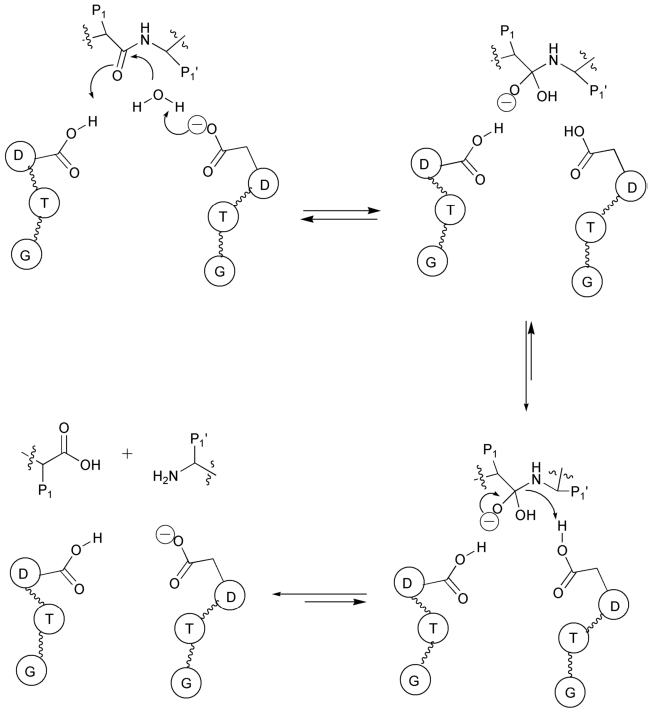


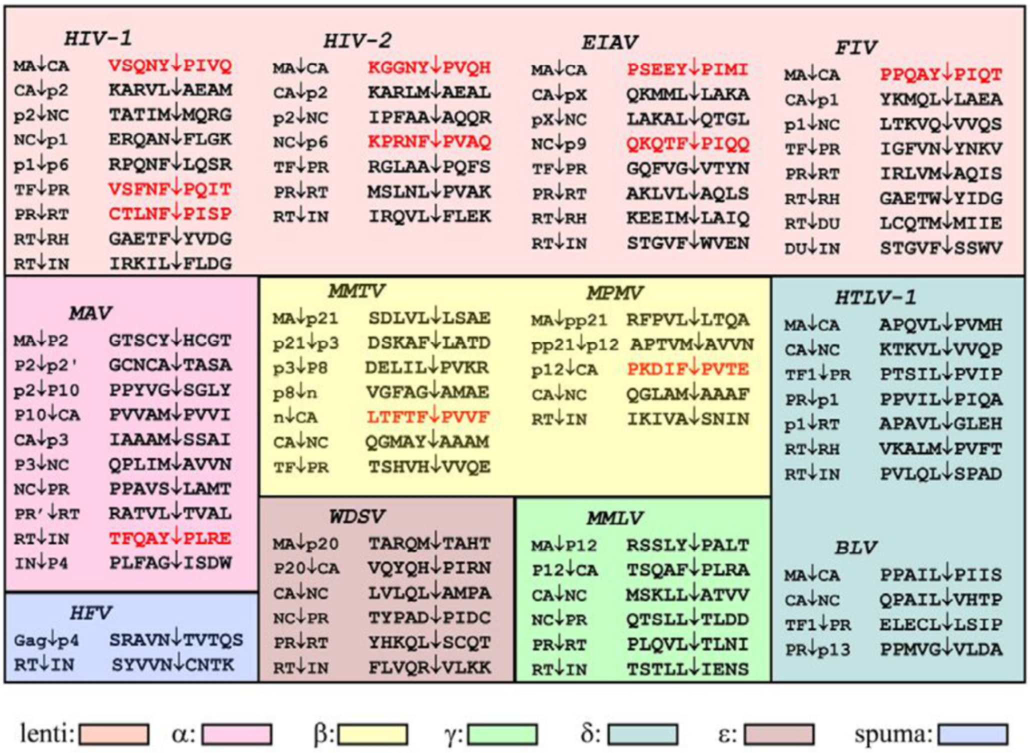
Figure 3. *General acid-base reaction mechanism for catalysis of APs.* Starting from the upper-left angle and following the reaction clockwise: binding of the substrate and nucleophilic attack of water; formation of the tetrahedral gem-diol intermediate; protonation of the nitrogen atom; formation of the products and release of the products. Adapted from Brik and Wong[73].

The significant roles of APs in human diseases led to massive efforts in the understanding of structure-function relationships of these enzymes. Together with the comparison of 3D structures, subsite specificity studies of numerous APs have provided in-depth knowledge of their catalytic mechanisms, which in turn contributed to the structure-based design and synthesis of a broad range of AP inhibitors for the treatment of many human diseases. The specificity subsites of APs are formed by hydrophobic residues surrounding the catalytic Asp dyad and by the residues in the flap-turn, as previously mentioned. The majority of the members of families A1 and A2 show preference for the cleavage of peptide bonds of at least eight residues (P4-P4' positions) in an extended conformation (the neighboring residues have their side chains projected in opposite directions). Despite the similar bond cleavage apparatus/catalytic mechanism, the substrate specificity and the enzymatic sites which define this specificity are much different in these proteases. In fact, while the specificity requirements of enzymes involved in general proteolysis (e.g., pepsin or cathepsin D) is usually nonstringent, enzymes involved in the regulation of physiological processes, such as renin, exhibit highly stringent substrate specificity[74,75].

Like the majority of members of family A1, pepsin is known to have the major specificity determinants at P1 and P1' subsites. This AP prefers to cleave after large hydrophobic residues, such as Phe and Leu (P1), whereas it hardly cleaves after His and Lys (unless they are next to Leu, Phe and a few others)[74,76,77]. The P1' position is less stringent exhibiting a preference for the residues Trp, Tyr, Phe and Val. The distal subsites are in general quite non-specific and can accommodate different types of residues, but have also been reported to play a role in the specificity of these enzymes. For instance, it was shown that positively charged residues at P3 position have detrimental effects in pepsin activity[77].

Contrastingly to the broad specificity of pepsin, renin displays a restricted selectivity for the amino acid sequence on either side of the peptide bond, tolerating only little variations to the sequence of its natural substrate (angiotensinogen). Noteworthy, all positions from P4 to P4' interact directly with the catalytic pocket of renin and have significant effects on its activity, as it was shown that the shortest peptide that this protease can cleave is the octapeptide Pro-Phe-His-Leu|Leu-Val-Tyr-Ser (| denotes the cleavage site)[78]. Moreover, although P1', P1, and P3 residues of renin substrates have been identified to be critical for its activity[75], it was also found that the His-Pro-Phe motif of angiotensinogen is a key determinant of the substrate specificity, by recruiting the scissile peptide bond to a favorable site for catalysis[79].

What is not so obvious is the substrate specificity of retroviral proteases, mainly because they do not have a particular consensus sequence substrate. Nevertheless, the analysis of similarities in amino acid sequences of a broad range of cleavage site sequences suggested their classification into two major groups: type 1 and type 2 (Figure 4)[80,81]. Type 1 cleavage sites are characterized by having an aromatic residue and Pro at P1 and P1' position, respectively, while type 2 sites have hydrophobic residues (excluding Pro) at both sides of the scissile bond[80]. The type 1 cleavage site is remarkable since, with the exception of pepsin, no other protease is known to act at the imino side of a Pro residue. The P2 and P2' positions are also critical in determining the type of cleavage site. The S2/S2' subsites are mostly hydrophobic and smaller than the S1/S1' or the S3/S3' binding sites and have been shown to be more specific, restricting the type and size of residues at P2/P2' in substrates or inhibitors relative to other binding pockets in the protease molecule; S1/S1' and S3/S3' subsites have a rather broad specificity due to their ability to accept residues of different types and sizes[66,82,83].



HIV-1

MA↓CA	VSQNY↓PIVQ
CA↓p2	KARVL↓AEAM
p2↓NC	TATIM↓MQRG
NC↓p1	ERQAN↓FLGK
p1↓p6	RPQNF↓LQSR
TF↓PR	VSFNF↓PQIT
PR↓RT	CTLNF↓PISP
RT↓RH	GAETF↓YVDG
RT↓IN	IRKIL↓FLDG

HIV-2

MA↓CA	KGGNY↓PVQH
CA↓p2	KARLM↓AEAL
p2↓NC	IPFAA↓AQQR
NC↓p6	KPRNF↓PVAQ
TF↓PR	RGLAA↓PQFS
PR↓RT	MSLNL↓PVAK
RT↓IN	IRQVL↓FLEK

EIAV

MA↓CA	PSEEY↓PIMI
CA↓pX	QKMML↓LAKA
pX↓NC	LAKAL↓QTGL
NC↓p9	QKQTF↓PIQQ
TF↓PR	GQFVG↓VTYN
PR↓RT	AKLVL↓AQLS
RT↓RH	KEEIM↓LAIQ
RT↓IN	STGVF↓WVEN

FIV

MA↓CA	PPQAY↓PIQT
CA↓p1	YKMQL↓LAEA
p1↓NC	LTKVQ↓VVQS
TF↓PR	IGFVN↓YNKV
PR↓RT	IRLVM↓AQIS
RT↓RH	GAETW↓YIDG
RT↓DU	LCQTM↓MIIE
DU↓IN	STGVF↓SSWV

MAV

MA↓P2	GTSCY↓HCGT
P2↓p2'	GCNCA↓TASA
p2↓P10	PPYVG↓SGLY
P10↓CA	PVVAM↓PVVI
CA↓p3	IAAAM↓SSAI
P3↓NC	QPLIM↓AVVN
NC↓PR	PPAVS↓LAMT
PR'↓RT	RATVL↓TVAL
RT↓IN	TFQAY↓PLRE
IN↓P4	PLFAG↓ISDW

HFV

Gag↓p4	SRAVN↓TVTQS
RT↓IN	SYVVN↓CNTK

MMTV

MA↓p21	SDLVL↓LSAE
p21↓p3	DSKAF↓LATD
p3↓P8	DELIL↓PVKR
p8↓n	VGFAG↓AMAE
n↓CA	LTFTF↓PVVF
CA↓NC	QGMAY↓AAAM
TF↓PR	TSHVH↓VVQE

MPMV

MA↓pp21	RFPVL↓LTQA
pp21↓p12	APTVM↓AVVN
p12↓CA	PKDIF↓PVTE
CA↓NC	QGLAM↓AAAF
RT↓IN	IKIVA↓SNIN

HTLV-1

MA↓CA	APQVL↓PVMH
CA↓NC	KTKVL↓VVQP
TF1↓PR	PTSIL↓PVIP
PR↓p1	PPVIL↓PIQA
p1↓RT	APAVL↓GLEH
RT↓RH	VKALM↓PVFT
RT↓IN	PVLQL↓SPAD

WDSV

MA↓p20	TARQM↓TAHT
P20↓CA	VQYQH↓PIRN
CA↓NC	LVLQL↓AMPA
NC↓PR	TYPAD↓PIDC
PR↓RT	YHKQL↓SCQT
RT↓IN	FLVQR↓VLKK

MMLV

MA↓P12	RSSLY↓PALT
P12↓CA	TSQAF↓PLRA
CA↓NC	MSKLL↓ATVV
NC↓PR	QTSLL↓TLDD
PR↓RT	PLQVL↓TLNI
RT↓IN	TSTLL↓IENS

BLV

MA↓CA	PPAIL↓PIIS
CA↓NC	QPAIL↓VHTP
TF1↓PR	ELECL↓LSIP
PR↓p13	PPMVG↓VLDA

lenti: ▭ α: ▭ β: ▭ γ: ▭ δ: ▭ ε: ▭ spuma: ▭

Figure 4. *Processing sites in retroviral Gag and Gag-Pro-Pol polyproteins.* The naturally occurring cleavage sites in retroviruses human immunodeficiency virus 1 (HIV-1), human immunodeficiency virus 2 (HIV-2), equine infectious anemia virus (EIAV), feline immunodeficiency virus (FIV), myeloblastosis associated virus (MAV), mouse mammary tumor virus (MMTV), Mason- Pfizer monkey virus (MPMV), human foamy virus (HFV), Walleye dermal sarcoma virus (WDSV), Moloney murine leukemia virus (MMLV), human T-lymphotropic virus 1 (HTLV-1) and bovine leukemia virus (BLV) are indicated by an arrow. Cleavages sites of type 1 are indicated in red. Gag and Gag-Pro-Pol regions: matrix (MA), capsid (CA), nucleocapsid (NC), transframe protein (TF), protease (PR), RNaseH (RH), integrase (IN) and dUTPase (DU); proteins and peptides with unidentified functions are abbreviated with the size of the protein in kDa (e.g., p12 is a protein having 12 kDa, while pp refers to phosphoprotein), as pX, or by the letter number. The type of virus and corresponding color is indicated on the bottom of figure. From Tözsér[66].

In type 1 cleavage sites, there is a preference for Asn at P2 and for β-branched hydrophobic residues (Val or Ile) at P2', while in type 2 cleavage sites the P2 position is typically β-branched and the P2' residue is Glu or Gln. Despite this overall classification in type 1 and type 2 cleavage sites, systematic specificity studies on HIV-1 PR revealed that the optimal enzyme-substrate interaction cannot be described residue by residue and that the specificity determinants are not confined to the P2-P2' region[81,82]. Additionally, many of the natural cleavage sites of retroviral proteases do not fit into this classification [e.g., they contain Pro after a nonaromatic residue or contain polar residues at P1 or P1' (Figure 4)], and therefore this classification might be considered as an oversimplification of a more complex pattern[66]. In fact, different studies have shown that the preference for a residue at a particular position in HIV-1 PR substrate is strongly dependent on the sequence context and conformation of the peptide substrate, including both neighboring residues at the same side and at the opposite side of the peptide backbone of the substrate[66,81,82]. These detailed specificity studies provided an explanation for the lack of a consensus substrate sequence for HIV-1 PR and for the majority of retroviral proteases and have also demonstrated that specificity towards nonviral protein substrates significantly differed from viral polyprotein cleavage sites: unlike in the Gag and Gag-Pol cleavage sites, cellular protein cleavage sites frequently contain charged residues, especially the Glu at P2'[84]. Moreover, although retropepsins are symmetrical dimers, no obvious symmetrical substrate preference has been observed for the specificity of HIV-1 PR. In fact, an asymmetrical arrangement adopted by the substrate peptides of HIV-1 PR has been suggested to be preferred over a particular amino acid sequence[51].

Given the pivotal role of many APs, they have been recognized as an important group of enzymes in scientific, medical research and biotechnology, some of them are already used as key drug targets. The advance of structural biology, high-speed parallel synthesis, computational chemistry, and drug development have greatly contributed to the discovery and optimization of several AP inhibitors[58]. Nearly all known APs are inhibited by pepstatin[85], a naturally occurring pentapeptide produced by *Streptomyces* strains, containing the unusual amino acid statine (Sta, (3S,4S)-4-amino-3-hydroxy-6-methylheptanoic acid). Pepstatin has been widely used in pharmaceutical industry as a model inhibitor. Other specific AP inhibitors include Diazoacetylnorleucinemethyl ester (DAN) and 1,2-epoxy-3-(p-nitrophenoxy)propane (EPNP), which inhibit most APs in the presence of cupric ions[86].

The recognition of the HIV-1 PR as a member of the AP family[87] and of its role in the maturation of HIV-1 has renewed the interest in this type of enzymes and on their inhibition, in order to arrest virus development[88,89]. In fact, the introduction of HIV-1 PR inhibitors represented a milestone on AIDS treatment. Since the first HIV-1 PR inhibitor approved by FDA in 1995, saquinavir, other nine inhibitors have become commercially available. These drugs are competitive inhibitors for the active site of the protease and all of the inhibitors are peptidomimetics, with the exception of tripanavir. Peptidomimetic inhibitors bind to the active site via an extensive network of hydrogen bonds, thereby mimicking the transition state of the substrate, but they are not cleaved due to its hydroxyethylene or hydroxylethylamine core[90]. By preventing the action of the protease, the viral maturation and replication process is blocked. Nevertheless, the lack of proofreading activity of reverse transcriptase results on the introduction of frequent mutations during reverse transcription of viral RNA to DNA. These mutations ultimately lead to several structural changes on HIV-1 PR and, therefore, limits the efficacy of currently used protease inhibitors as a result of broad cross resistance[91,92]. However, the large amount of kinetic and crystallographic studies has increased the knowledge on enzyme function, structure and catalytic mechanism which led to a better understanding of how drug-resistance mutations exert their effects at a molecular level. These insights are valuable for the design of new drugs and therapeutic strategies to combat drug resistance to AIDS as well as to combat other virus-related human/mammalian malignancies.

Another example of an AP inhibitor approved by the FDA as a therapeutic agent is Aliskiren, a non-peptidic inhibitor of renin which has been used for the treatment of hypertension[93]. Renin participates in the rate-limiting step of the renin-angiotensin system (RAS), by hydrolyzing angiotensinogen into angiotensin, which is further converted into angiotensin II by the angiotensin-converting enzyme (ACE). Because of renin specificity, their inhibitors are potent anti-hypertensive agents similar in action to ACE inhibitors. Furthermore, attempts to develop new renin inhibitors have been hampered by the peptidic character of the new molecules, which confers low stability and poor oral bioavailability in humans associated with higher production costs[24].

Notably, the achievement of selective inhibition of renin and HIV-1 PR has provided an unambiguous validation that AP inhibitors can be successful drugs for improving human/mammalian condition. Despite this progress, many other inhibitors have failed on selectively inhibiting the targeted AP. In fact, several inhibitors of BACE1, cathepsin D or plasmepsins developed to date are also potent inhibitors of their counterparts, and thus better bioavailability, specificity and potency are needed for maximizing the inhibition of the target enzyme without causing toxicity and/or undesired side effects[23,24].

Evolutionary relationships of aspartic proteases

Very limited similarities are found in the amino acid sequences of the two homologous lobes of pepsin enzymes. Strikingly, the retroviral protease monomer exhibits nearly the same three-dimensional organization of a single lobe of pepsins. The most conserved parts correspond to those regions in the N- and C-terminal domains of pepsin-like enzymes that are related by the interdomain dyad and consist of a combination of secondary structural elements that form the psi-loop motif at the active site of APs. Regardless of the differences between the interdomain antiparallel β-sheet of pepsin-like APs and the intersubunit β-sheet of retropepsin enzymes in number, arrangement and directionally of strands, a detailed topological analysis of these structures revealed that A1 and A2 family members share a direct structural relationship[94]. Experimental studies carried out to understand the structural and functional relationships among the two families of APs, have shown latter on that dimers of the N-terminal lobe of porcine pepsinogen can express some catalytic activity, thereby confirming that the interface of the lobes was conserved throughout evolution[95].

Established that the aspartic dyad is located at the interface region and that the viral subunits are structurally similar to the N-terminal lobe of pepsin-like enzymes, it became clear that the two families are indeed evolutionary related. In fact, it is well-accepted that pepsin-like enzymes have likely evolved from primordial single lobed APs as a result of a gene duplication and fusion event[42,95,96]. However, the evolutionary relatedness between eukaryotic and retroviral APs has been the subject of controversy among experts for many years in regard to the nature of the common ancestor. The lack of clear evidences (until recently, as will be discussed) for the presence of both bilobal and single lobed APs in prokaryotes have always suggested that this gene duplication event would have occurred after divergence of bacteria and eukaryotes. In line with this, two main theories have been advanced on how these enzymes could have evolved. The first (and most accepted) proposes that retroviral proteases may represent a direct precursor of eukaryotic bilobal pepsins and, therefore, the homodimeric APs encoded by retroviruses, pararetroviruses and retrotransposons would correspond to the most ancestral state of these enzymes (Figure 5); whereas the second theory suggests that retroviral proteases may have evolved from the acquisition of a pepsin-like gene followed by one or more deletion events, similar to certain viral oncogenes that are deletion products of cellular proto-oncogenes[94].

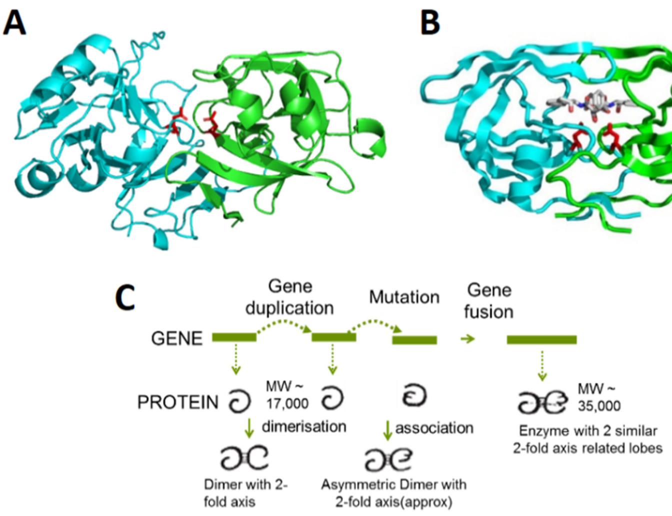


Figure 5. *Evolutionary theory of gene duplication and fusion in the origin of pepsin-like enzymes.* Despite the structural differences and the low sequence identity, it is believed that the (**A**) pepsin and (**B**) retropepsin families are evolutionarily related since, in both folds, the cleavage site loops are homologous, the Asp dyad is located at an interface region, and the viral subunits are structurally similar to the N-terminal lobes of pepsin-like family members. (**C**) According to these observations, pepsin-like enzymes have been suggested to have derived from the duplication and fusion of a gene encoding a single-lobed AP[96].

Interestingly, this later hypothesis that retroviral proteases may constitute a derived state, and likely be the result of a more recent gene transfer, has been recently strengthened by the identification of genes encoding single-lobed (retroviral-like) APs in eukaryotes, as well as in protozoans and prokaryotes, that are not embedded within endogenous retroviral elements[97–100].

The *Saccharomyces cerevisiae* DNA-damage inducible protein 1 (Ddi1) was the first identified retroviral-like AP in eukaryotes[101]. This protein was primarily identified as a suppressor of a temperature sensitive mutant of the *PDS1* gene, which encodes Pds1p, a key regulator of the cell cycle whose ubiquitin-dependent proteolysis initiates anaphase in budding yeast. Ddi1 is conserved in all eukaryotes for which sequence information is available, including the early-branching protozoans *Leishmania* and *Plasmodium*, suggesting that this protein may have an ancient, critical function. Bioinformatics methods revealed that Ddi1 and its orthologs in *Drosophila* and *Arabidopsis* share a common domain architecture, with an amino-terminal ubiquitin domain, a central protease domain containing the typical Asp-Ser/Thr-Gly signature, and a carboxy-terminal ubiquitin associated domain[97]. In fact, although not proven yet to be a protease, the crystal structure of this domain of Ddi1 shows that it is a dimer with a fold similar to that of the retroviral proteases, thereby suggesting that Ddi1 may function proteolytically during regulated protein turnover in the cell[102]. More recently, the homologue to the Ddi1 protein from *Leishmania major* was shown to be an active AP with preferential substrate selectivity for the retropepsin family substrates, an optimal activity in acidic conditions and an inhibition of BSA degradation under the presence of DAN, pepstatin and nelfinavir[103]. Molecular modeling of the retroviral domain of the Ddi1-like *Leishmania* protein revealed a dimer interface encompassing two Asp-Ser-Gly-Ala amino acid sequence motifs, in an almost identical geometry to the displayed by the homologous retroviral protease domain of *Saccharomyces* Ddi1 protein[103].

Skin aspartic protease (SASPase) is another example of a retropepsin that was found to be specifically expressed in the stratum granulosum (SG) of the human epidermis[100]. Using high-throughput *in situ* hybridization screening, a mouse homolog of SASPase was also identified as an SG-expressing gene[104]. This retroviral-like AP was only found in mammals, and immunoblotting of human and mouse epidermal extracts revealed the expression of two forms of the enzyme: the 28 and 14 kDa forms in human and the 32 and 15 kDa forms in mouse. Although human and mouse SASPase are predicted to have one transmembrane domain, the full-length protein was never detected in epidermal lysates of both species, and therefore, the function of the transmembrane domain remains to be elucidated. Nevertheless, it was shown that both the human (28 kDa) and mouse (32 kDa) recombinant SASPase forms undergo auto-activation processing *in vitro*, similar to other retroviral proteases such as HIV-1 PR, and that this cleavage event generates a 14 kDa (human)/ 15 kDa (mouse) derived protease domain. Also similar to HIV-1 PR, mouse SASPase displayed an optimal activity pH of 5.77, which corresponds to the pH of the upper surface of the epidermis. Human SASPase was also shown to be insensitive to pepstatin while auto-activation was inhibited by the HIV-1 PR inhibitor indinavir[100]. More recently, it was observed that SASPase activity is indispensable for processing profilaggrin and maintaining the texture and hydration of the stratum corneum, thereby preventing fine wrinkle formation[100,105].

Taken together, the recent discovery of Ddi1-like protein from *Leishmania major* and SASPase with their unique properties of auto-activation (in the case of SASPase) and subsequent dimerization to form an active enzyme, provide new arguments in favor of a homodimeric eukaryotic AP being at the origin of retroviral proteases via horizontal transfer from a host at an early stage of eukaryotic evolution, and not the other way around as mostly accepted[100,103].

Solving the crystal structure of the protease encoded by Xenotropic Murine Leukemia virus–related virus (XMRV) by Li and colleagues[106], also strengthens this link between retropepsin-like homologues found in eukaryotes with those found in retrovirus. In fact, despite its overall structural similarity to other retropepsins, the XMRV PR displays a novel type of dimeric interface. Some structural features of XMRV PR, specially the longer N and C termini forming hairpins, clearly defines a more closely resemblance in structure with pepsin-like enzymes[106]. In fact, instead of the typical interdigitation of C- and N-terminal β strands between subunits of retropepsins, only the C-terminal β-strands of XMRV PR form the dimer interface. Accordingly, both C terminus are topologically and structurally equivalent to the corresponding C-terminal loops of each domain of pepsin-like enzymes which form a six-stranded interface, with the difference that in XMRV PR this interface is four-stranded[106]. Together with other unique features, these observations suggest a closer resemblance of XMRV PR over other retroviral proteases to the putative common ancestor of pepsin and retropepsin-like enzymes, thereby supporting the theory that single-chain pepsin-like APs arose by gene duplication and divergence[106].

As previously mentioned, APs were always assumed to be restricted to eukaryotes and viruses. However, the observations of Rawlings and Bateman[16] started changing this paradigm by proposing that the pepsin family of peptidases was not just confined to eukaryotes, but also present in bacteria. Through genomic sequence analysis, proteins bearing the characteristic hallmark features of the pepsin family were found in seven genomes of bacteria that belong to the class γ-proteobacteria: *Colwellia psychrerythraea, Marinomonas* sp. MWYL1*, Shewanella amazonensis, Shewanella denitrificans, Shewanella loihica, Shewanella sediminis* and *Sinorhizobium medicae*. These bacterial homologues were predicted to be structurally similar to pepsin, consisting of two lobes each of which bearing one active site Asp. The experimentally validation for these observations was latter given by Simões and colleagues, which have expressed in *E. coli* the AP gene from *Shewanella amazonensis*, named shewasin A[17]. This pepsin-like enzyme exhibits activity at acidic pH against a well-documented AP substrate, preferentially between hydrophobic amino acids, and is inhibited by pepstatin. Contrary to its closest eukaryotic homologues, shewasin A do not require a propeptide or signaling sequence for correct folding or secretion to the periplasm. The detailed biochemical characterization of recombinant shewasin A together with its positioning close to the divergence of the two subfamilies on the phylogenetic tree, has clearly demonstrated that this bacterial AP is strongly reminiscent of its eukaryotic counterparts[16].

These important observations have raised the discussion on the evolutionary relationships between bacterial and eukaryotic pepsin-like APs, by suggesting that bi-lobal pepsin-like proteases may have evolved from a primordial homodimeric AP before divergence between eukaryotes and prokaryotes, through the proposed duplication and fusion events[42]. In line with this view, single lobed APs would also have to be present in bacterial organisms (raising again the discussion whether retroviral retropepsins might represent a derived state). In this regard, the identification of two bacterial proteins SpoIIGA[98] and Perp[99], exhibiting the catalytic determinants resembling those of the retropepsin type, have provided the first evidences to support this hypothesis. SpoIIGA, a novel type of AP, has been identified in *Bacillus* species as required for processing of pro-σ_E to σ_E factor in endospore formation of *Bacillus subtilis*[98]. According to the authors, this membrane-embedded protein shares structural features with the HIV-1 PR in the C terminus region and this domain was assumed to be active as a dimer because of the presence of a single Asp-Ser-Gly motif. However, as previously argued for other proteins[16], this alone appears insufficient to include SpoIIGA in the retropepsin family because the same motif occurs around the active site Asp residue of the serine-type peptidase subtilisin. Furthermore, since no enzymatic characterization is yet available, it has been highly controversial to consider the existence of one AP catalytic sequence motif by itself as a strong evidence to prove the existence retroviral proteases in prokaryotes[98]. For this reason, SpoIIGA has been accepted in the *MEROPS* database as a new family of APs (A36).

Almost at the same time, another protease named PerP, was identified by Chen and colleagues and was also suggested to belong to the retropepsin family[99]. This periplasmic protease removes a C-terminal peptide producing a truncated form of PodJ (a polar factor that recruits proteins required for polar organelle biogenesis to the correct cell pole), important in the sessile phase of the cell cycle of *Caulobacter crescentus* to recruit components for stalk assembly[99]. PerP was identified as containing a putative signal sequence or membrane anchor at its N terminus and a conserved AP catalytic motif Leu-Val-Asp-Thr-Gly-Ala in its periplasmic domain[99]. Even though no reference was made regarding its classification and no enzymatic characterization has been provided, the existence of a single catalytic aspartate residue assumed to be the active site residue, prompted the classification of PerP on *MEROPS* database as a novel family of APs – family A32 (Clan AA).

Overall, due to the absence of characterization assays of SpoIIGA and PerP proteins, further investigations are still required to identify a functional retropepsin in bacteria and thus provide the unequivocal evidences that the hypothetical gene duplication and fusion events (that may have given rise to bi-lobal pepsin-like enzymes), have indeed preceded the most recent common ancestor of prokaryotes and eukaryotes.

1.2. *Rickettsiae* and Rickettsioses

Rickettsiae are a diverse group of organisms, some of which causing mild to severe infection diseases in humans and animals, such as Rocky Mountain spotted fever (RMSF) and Mediterranean spotted fever (MSF). They are genetically related α-proteobacteria with fascinating obligatory intracellular lifestyle and are maintained in nature through a cycle involving reservoir in mammals and arthropod vectors[107]. The impact of these pathogens in public health is largely unmeasured, but assumed to be fairly high worldwide. In fact, over the last years there has been a growing concern about rickettsial diseases and their impact on global health as members of the genus *Rickettsia* have been identified together with other bacteria as emerging/re-emerging pathogens, responsible for the majority of infectious diseases on the last 60 years[108,109]. *Rickettsia* are life-threatening pathogens not only due to their highly virulent properties but also due to a set of unique biological characteristics, such as environmental stability, aerosol transmission, persistence in infected hosts and low infectious dose, which makes them a potential powerful biological weapon[110,111].

Being rickettsiae the causative agents of the some of the most severe human infections, with sophisticated and highly effective pathogenic strategies, they have been the target of many studies. As a result of this, important progresses in research into genomics and pathogenesis of these bacteria as well as regarding the immune responses to these microorganisms have been made. This section presents a revision on the current knowledge of the pathogenicity developed by *Rickettsia* throughout evolution.

Bacteriology and Epidemiology

Bacteriology

Rickettsiae represents a large and metabolically diverse group of gram-negative bacteria that have the capacity to infect and replicate in the cytosol and occasionally in the nucleus of vertebrate cells (e.g., endothelium, vascular smooth muscle, and macrophages) and invertebrate cells (e.g., hemocytes and salivary gland epithelium)[112,113]. These small (0.8–2 μm x 0.3–0.5 μm), non-motile, short rod-shaped, coccobacillary gram-negative bacteria divide by transverse binary fission and stains poorly with conventional Gram techniques, but retain basic fuschin when stained using the Gimenez method[110,114].

Epidemiology

Rickettsioses represent some of the oldest recognized pathologies transmissible from animal to humans, many of them causing mild to fatal diseases. Despite being longstanding diseases, it was not until the early part of the 20th century that the first report was given by Howard T. Ricketts for the involvement of a transmissible *Rickettsia* causing RMSF[115]. The investigations carried out in the Bitterroot Valley of Western Montana have been stimulated by the spring seasonality and high incidence of this severe disease (known locally as "spotted fever"), with a case-fatality rate as high as 64%. Throughout history, rickettsioses have played a significant role on Western civilization causing more deaths than all the wars combined and accounting for 22.8% of deadly emerging infectious diseases in humans, in the absence of timely and appropriate antibiotic treatment (Table 3)[109,116].

Epidemic typhus is one of the most dangerous arthropod-borne disease affecting mankind with a widespread occurrence and mortality rates ranging from 10 to 60%[117]. From the 15th through 20th centuries, epidemic typhus has killed millions of people, particularly during or immediately after World Wars I and II, thus affecting the course of European history[118]. Transmission of the agent by the human body louse was proven by Nicolle in 1909[119,120], and in 1916 Da Rocha-Lima proved that *Rickettsia prowazekii* was the etiologic agent[121]. Epidemic typhus has re-emerged fairly recently in louse-infested populations, particularly in developing countries with a context of socio-political instability, famine, civil wars or natural disasters[108]. In this perspective, the threat of louse-borne typhus is still real and because *R. prowazekii* is a potential bioterrorism weapon with an infectious dose lower than 10 organisms, this pathogen has been classified as a category B NIAID (National Institute of Allergy and Infectious Diseases) Priority Pathogen, whereas the other rickettsiae fall in category C[108,113].

Like epidemic typhus, RMSF is also a highly virulent human infection, with significant morbidity and mortality, and potentially fatal even in healthy young individuals. RMSF is caused by *Rickettsia rickettsii*, a member of the spotted fever group (SFG) of the genus *Rickettsia.* The principal vectors of RMSF in the United States are *Dermacentor variabilis* and *Dermacentor andersoni*, which are most active during late spring and summer, when RMSF peaks. Cases of RMSF have been reported in 48 states, but 64% of these cases were reported from only five states: North Carolina, Oklahoma, Arkansas, Tennessee and Missouri[122]. *R. prowazekii* is capable of surviving within infected individuals for the lifetime of the host and, under extreme stress conditions, these latent bacteria can become active and cause a relapsing form of epidemic typhus known as Brill-Zinsser disease[123].

Table 3. *Rickettsial diseases in humans.* Adapted from Walker and Ismail[108].

Disease	Organism	Arthropod vector	Life cycle	Geographic area	Symptoms of fever	Mortality rate *
Rocky Mountain spotted fever	*R. rickettsii*	*Dermacentor variabilis, Dermacentor andersoni, Rhipicephalus sanguineus, Amblyomma cajennense* and *Amblyomma aureolarum*	Transovarial in ticks and rodent	Western hemisphere	Yes	High
Boutonneuse fever	*R. conorii*	*Rhipicephalus sanguineus* and *Rhipicephalus pumilio*	Transovarial in ticks	Southern Europe, Africa and southern Asia	Yes	Mild to moderate
African Tick Bite fever	*R. africae*	*Amblyomma hebraeum* and *Amblyomma variegatum*	Transovarial in ticks	Africa and the West Indies	Yes	None reported
Maculatum disease	*R. parkeri*	*Amblyomma maculatum* and *Amblyomma triste*	Ticks	Western hemisphere	Yes	None reported
Flea-borne spotted fever	*R. felis*	*Ctenocephalides felis*	Transovarial in the cat flea	Worldwide	Yes	None reported
Murine typhus	*R. typhi*	*Xenopsylla cheopis* and *Ctenocephalides felis*	Rat-flea for *Xenopsylla cheopis* and opossum flea for *C. felis*	Worldwide	Yes	Low
Epidemic typhus	*R. prowazekii*	*Pediculus humanus humanus*	Human louse	Worldwide	Yes	High
Epidemic typhus	*R. prowazekii*	*Fleas and lice of flying squirrels* and *Glaucomys volans volans*	Flying-squirrel flea and louse ectoparasite	United States	Yes	Low
Rickettsialpox	*R. akari*	*Liponyssoides sanguinus*	Transovarial in mites	Worldwide	Yes	None reported

*High mortality is >15%; moderate mortality is 7–15%; mild-to-moderate mortality is 2–7% and low mortality is ≤1%.

Transmission of rickettsial diseases by previously unknown and unexpected arthropod vectors further demonstrates the ability of the pathogen to adapt to new ecological niches and maintain virulence[124]. Other factors that might be contributing to the emergence and global spread of rickettsioses include the increasing proximity of human and animal populations which is the result of the human population growth and their mobility for socioeconomic, cultural and recreational purposes[125,126].

Even though more than one century has passed since the first description of RMSF, the majority of the newly identified species and subspecies of rickettsiae have been described as emerging pathogens to humans only over the past 30 years[109]. Owing to the improved diagnostic methods and increasing interest, rickettsiae are now recognized in all parts of the world in endemic foci with sporadic and often seasonal outbreaks: *R. japonica* in Japan and Korea; *R. honei* in Australia and Southeast Asia; *R. africae* throughout sub-Saharan Africa and in the French West Indies; *R. felis* globally; *R. sibirica* mongolo-timonae strain in Asia, Europe, and Africa; *R. parkeri* in North and South America; *R. heilongjiangensis* in northeastern Asia; *R. aeschlimannii* in Africa; and *R. helvetica* and *R. canadensis* also suspected to be human pathogens[110]. In recent years, a number of new *Rickettsia* spp. has been discovered in Europe, some of which have been shown to be pathogenic to humans. MSF (also known as boutonneuse fever), the most common and well characterized rickettsial human disease in the Mediterranean region, is a tick-borne rickettsial disease caused by *Rickettsia conorii* and transmitted by *Rhipicephalus sanguineus* ticks. In general, MSF is considered to be a milder rickettsial disease with a lower mortality rate than RMSF[127], but the increasing geographic distribution (Southern Europe, North and West Africa, India, Pakistan, Israel, Russia, Georgia, and Ukraine) and severity have raised important concerns about human infections caused by *R. conorii*[128]. In 1997 in Beja, a Portuguese Southern district with climatic conditions favorable to *R. sanguineus*, the mortality rate in hospitalized patients with MSF was the highest ever obtained since 1994 (32.3%)[129,130]. Including *R. conorii*, eight tick-borne species or subspecies within SFG have been reported as emerging pathogens in Eastern and Southern Europe[131]. Moreover, of the non-tick-borne species, *R. felis*, associated with cat fleas, is also an emerging human pathogen[132,133] and the mite-transmitted *R. akari*, the agent of Rickettsialpox, is known to be prevalent in Europe too[131].

Phylogenomics

<u>*Taxonomy and Phylogeny*</u>

The genus *Rickettsia* is included in the bacterial tribe rickettsiae within the family Rickettsiaceae in the order *Rickettsiales*, a highly diverse collection of early-branching lineage of the α-proteobacteria. Some members of the genus *Rickettsia* are recognized human pathogens, while others should preferentially be considered as species or strains of unknown pathogenicity than as nonpathogenic, particularly when associated with arthropods able to bite humans[134].

The proliferation of named species over the past three decades has generated controversy among rickettsiologists regarding the appropriate taxonomy of *Rickettsia* spp. As for other prokaryotes, traditional phylogenetic studies of *Rickettsia* were based on the comparison of morphological, ecological, epidemiological and clinical characteristics to differentiate and describe new bacterial species. Members of this genus have classically been separated into the SFG and those from the typhus group (TG), based on the differences in the diseases they cause, differences in their antigenicity to lipopolysaccharide and outer-membrane proteins (rOmpA and rOmpB), and on the ability to promote intracellular actin-based motility[135–137]. However, phylogenetic relationships based on these criteria were highly unreliable and some *Rickettsia* spp. did not fit well within this grouping. Moreover, given the strictly intracellular life of rickettsiae, and thus the few phenotypic characters expressed, other traditional identification methods used in bacteriology were not applicable to *Rickettsia* spp. The advent of molecular methods has deeply modified the definition of ''*Rickettsia*'' and has allowed new taxonomic and phylogenetic inferences. The considerable sequencing efforts over the last 20 years have culminated in the annotation of 63 complete genomes of *Rickettsiales* validated species, many of which cause diseases in humans and animals[109]. Further extensive work from various research groups, primarily using comparative analysis of gene sequences, have significantly contributed for a reliable estimation of evolutionary relationships of species of the order *Rickettsiales* and for the identification of relationships between genotype and phenotype, one of the major goals of the genomics era[138–140]. The 16S rRNA gene (rrs) was the first gene used on comprehensive phylogenetic studies of the *Rickettsiales*, which have resulted on a contemporary classification that differs greatly from the traditional classification scheme[141]. Comparative analysis based on this gene sequence have shown that several of the bacteria classified in the order *Rickettsiales*, like *Rickettsiella grylli, Coxiella burnetii* and *Eperythrozoon* spp., do not belong to the α-proteobacteria subclass[142]. Additionally, many bacteria belonging to the order *Rickettsiales* have been reclassified into the three major families *Holosporaceae, Anaplasmataceae* and *Rickettsiaceae*, with the latter separated into two recognized genera: *Rickettsia* and *Orientia*[136,143–145].

Nevertheless, the 16S rRNA gene sequences were shown to be highly conserved within the genus *Rickettsia* and, therefore, significant inferences about intragenus phylogeny were not possible although these have confirmed the evolutionary unity of the genus[146]. As progressively more genes have been sequenced, molecular methods involving comparison of multiple genes have been developed, but it was only in 2003 that the first widely recognized molecular criteria for the speciation of *Rickettsiae* was published[147]. According to these genetic guidelines, a new isolate is classified as a new *Rickettsia* sp. when it has no more than one value above the following nucleotide sequence similarities by comparison with any validated *Rickettsia* spp.: ≥99.8 and ≥99.9% for rrs and gltA genes, respectively, and when amplifiable ≥98.8, ≥99.2, and ≥99.3% for rOmpA, rOmpB and gene D, respectively[147]. Since then, other multigenic approaches have been updated to include genes such as atpA, recA, virB4, dnaA, dnaK, rrl, combined with other characteristics such as DNA-DNA homology, G+C content, and single-nucleotide polymorphisms and multi-spacer typing analysis[142]. The use of these genetic criteria to the available rickettsial genome sequences consistently supported the revision of the long-standing classification of *Rickettsia* into either the TG or SFG. According to this revision, the TG is represented by only two species, the highly pathogenic and insect-associated *R. prowazekii* and *R. typhi*, which are the etiological agents of epidemic and murine (or endemic) typhus, respectively. In contrast, the SFG exhibits a marked expansion, comprising all rickettsial tick-borne virulent species, some of which being the causative agents of well-known tick-borne diseases, such as RMSF (*R. rickettsii*) and MSF (*R. conorii*)[148]. Even before gene sequencing, numerous studies based on molecular data have already shown that *R. canadensis* and *R. bellii* were the most divergent species within rickettsiae, clearly suggesting that they did not belong to any of the traditional groups but rather to a basal lineage[142]. Subsequent DNA sequencing revealed that these two species of unknown virulence possess larger genomes than the other rickettsial species sequenced so far and exhibit little colinearity with any of them. Further analysis of these genomes also suggested that these species may have retained several ancestral features lost in other lineages in the course of evolution, supporting the creation of the ancestral group (AG)[149]. Recent phylogenomic analysis have also revealed a distinct lineage that shares immediate ancestry with the members of the SFG, which has been named the transitional group (TRG)[150]. This sister clade of SFG includes the species *R. felis*, *R. akari* and *R. australis*, as well as symbionts of wasps (*Liposcelis* spp.), booklice (*Neochrysocharis* spp.)[113,151] and the mite *Ornithonyssus bacoti*[152]. One particular feature of this group is the tendency for its members to be associated with non-tick arthropod hosts[143].

Regardless all the aforementioned advances in serotyping and molecular genotyping of rickettsial species isolated from defined geographic locations, this taxonomic classification of *Rickettsia* into four groups (Figure 6) is still not consensual and alternative phylogenomic classifications have been proposed with the inclusion of several other distinct genetic groups[143,153]. The major issue is still the lack of universal consensus on the criteria that should be used for the designation of species, remaining unclear whether many of the new isolates described in recent years should be classified as new species or even subspecies, as they vary much less from one another than the species of other bacterial genera [110,154,155].

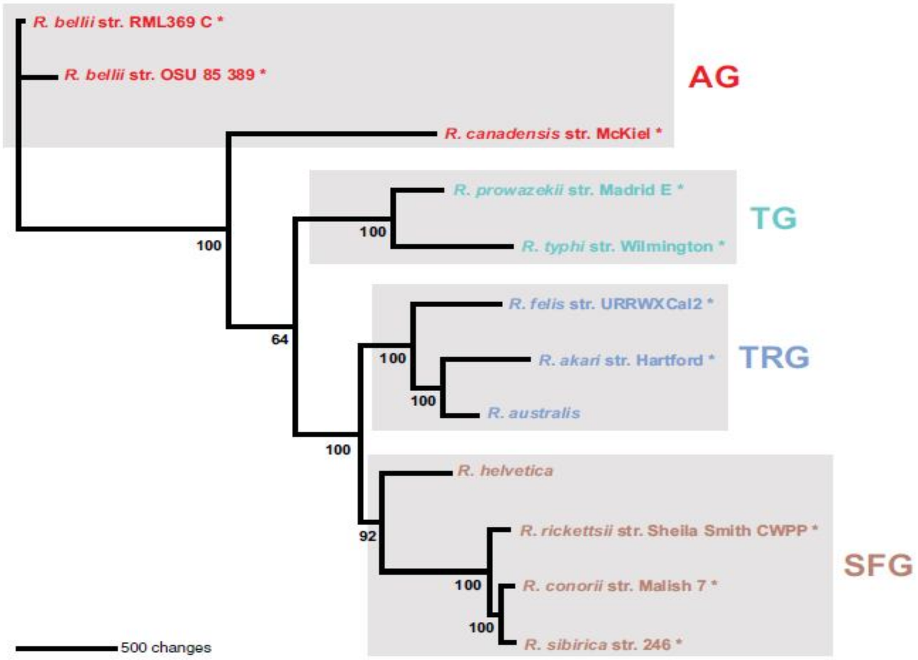

R. bellii str. RML369 C *
R. bellii str. OSU 85 389 *
R. canadensis str. McKiel *
AG
R. prowazekii str. Madrid E *
R. typhi str. Wilmington *
TG
R. felis str. URRWXCal2 *
R. akari str. Hartford *
R. australis
TRG
R. helvetica
R. rickettsii str. Sheila Smith CWPP *
R. conorii str. Malish 7 *
R. sibirica str. 246 *
SFG
100
100
64
100
100
100
92
100
100
500 changes

Figure 6. *Phylogenetic tree classification of Rickettsia spp. according Gillespie and colleagues*[143]. Phylogenomic analysis supported the reclassification of *Rickettsia* spp. into four groups: ancestral group (AG), typhus group (TG), transitional group (TRG) and spotted fever group (SFG).

Important evolutionary inferences have also emerged by comparing the sequences of rickettsial spp. and mitochondrial genes[139]. These phylogenetic studies have revealed that *Rickettsia* are more closely related to mitochondria than any other bacteria with a sequenced genome, with speculation that they evolved from a common ancestor[139]. Therefore, an important link has been established between *Rickettsiales* and the eubacterial ancestor of the mitochondria, although the placement of the latter within the eubacterial tree is still a subject of controversy[139,156]. As a matter of fact, despite several phylogenetic inferences have placed the mitochondrial ancestor within or basal to the *Rickettsiales*, and some specifically within the *Rickettsia*ceae[157], the most robust and accepted analysis placed the mitochondria as a sister taxon to *Rickettsiales* and *Anaplasmataceae*[156].

Genomics

Genomic research allied with studies on bacterial pathogenesis have uncovered noteworthy aspects of pathogen biology, such as the three main forces that shape the evolution of bacterial pathogens: gene gain, gene loss and gene change[158]. In this respect, the ever-increasing number of rickettsial sequenced genomes highlight unique characteristics among bacterial genomes, becoming an excellent model to investigate the process of reductive evolution[159]. In fact, while the most common prokaryotic genomes remain about the same size despite the acquisition of new genes over time by lateral gene transfer (gene gain must be balanced by gene loss), obligate intracellular bacteria such as *Chlamydia*, *Ehrlichia*, *Mycoplasma, Spirochaetes*, and *Rickettsia* have much more reduced genome sizes compared to their nearest free-living relatives. These organisms have undergone considerable genome downsizing as the result of the degradation and reduction of originally ancestral non-pathogen genomes, which invariably accompanied the adaptation to parasitic/symbiotic lifestyles[139]. In addition to the unusual G+C content of approximately 30%[142], investigations on a variety of hosts including non-hematophagous insects, amoebae and leeches, have revealed that *Rickettsia* genomes present substantial inter-species variations in size (1.1 Mb for the TG, 1.2-1.4 Mb for the SFG and 1.5 Mb for *R. bellii*) and gene content (about 900–1500 genes)[160,161]. These studies provide a strong indication that the rickettsial ancestral initiated intracellular parasitism in unicellular eukaryotes like amoebae and later adapted to multicellular eukaryotes[142]. The specialization to multicellular eukaryotes and latter to distinct arthropod hosts has been proposed to coincide with the beginning of rickettsial genome reduction and diversification. The fact that the obligate amoebal symbiont related to chlamydia has a large genome (2.4 Mb) compared to the related obligate intracellular human/animal pathogens (~1 Mb), exemplifies the importance of this type of host transition on the genome size[159,162].

The reductive evolution of *Rickettsia* genomes is mainly justified by the presence of orthologous genes in the host cells that compensate for the function of those rickettsial genes that have been discarded; a particular example are genes necessary for metabolite synthesis given the ability of bacteria to import proteins or metabolite products of the host genes[159,163]. Eventually, the replacement of many biosynthetic pathways present in free-living bacteria by transport systems in *Rickettsia* have resulted in a complete dependence upon the host cell for survival[163]. Also, it has been shown that the rate of sequence divergence, gene loss and genome rearrangements are tremendously variable throughout the various *Rickettsia* lineages, reflecting the intricate effects of specialization to distinct arthropod hosts as well as crucial alterations of the gene repertoire, including the amplification of mobile genes and the losses of DNA repair genes[159].

Horizontal gene transfer is a common event between prokaryotic organisms. Nevertheless, likely due to their strictly intracellular life cycle, rickettsiae minimize their exposure to horizontally transferred DNA, either with bacteriophages and transposons or with other species of bacteria, thereby exhibiting a low number of gene transfers and genome rearrangement[142]. Without exposure to such genetic parasites, there are no benefits of having a high chromosomal deletion rate. As a result, even under the ongoing process of genome reduction, the high number of split genes in these taxa reflect a reduction in the overall rates of chromosomal deletion[164]. Indeed, while in most bacteria the noncoding DNA or pseudogenes represents 10% of the genome, in the case of some rickettsial species this value can reach 24% (*R. prowazekii*)[165]. The few evidences for lateral gene transfer in *Rickettsia* have been provided by genome analyses through the identification of a large fraction of mobile genetic elements, including plasmids[142,160]. The presence of a conjugative plasmid in an intracellular bacterium has been first found on *R. felis*, suggesting that conjugation could play a role in the evolution of rickettsial genomes[150]. Plasmids have since been detected in *R. helvetica*, *R. peacockii* and *R. massiliae* along with a number of non-validated species[166].

In conclusion, the relatively low rate of lateral gene transfer and the continuous gene loss have resulted in highly conserved genomes exhibiting similar gene synteny and content, which has been related with a higher virulence of some *Rickettsia* spp.[136,159,162,163]. The development of new tools for comparative genomics has been critical to unveil many core genes encoding potential bacterial virulence factors, thereby providing important insights into the role of many proteins in pathogenesis[139,160,167].

Pathogenesis

Understanding the pathogenic steps of rickettsioses is essential for innovative interventions to halt disease progression. In this regard, several virulent factors have been identified over the last decades aided by the development of valuable tools in genomics and proteomics fields. The concept pathogenesis comprises three components: the sequence of events from transmission until immune clearance of the agent; the host–pathogen interaction from the cellular level to the whole patient; and the pathogenic mechanisms of cellular and tissue injury. In general, *Rickettsia* pathogenesis involves the following steps: transmission, entry in the organism, initial spread to other organs beyond the point of entry, adherence to and invasion of target cell, survival within the host, which implies evasion of the host defenses and adaptation to the host environment, and extension of the niche through modulation of host biology, multiplication and survival[168–171].

Life cycle

Transmission vectors

As *Anaplasma* and *Ehrlichia*, *Rickettsia* have arthropod hosts (e.g., ticks, mites, fleas or lice) which serve as the biologic vector that transmits the pathogens to animals and humans. The association of rickettsial species with obligate blood-sucking arthropods denotes the highly adapted end-product of years of biologic evolution. In light of these associations, which are characterized by efficient multiplication, long-term maintenance, transstadial and transovarial transmission, and ecologic and extensive geographic distribution, one can explain the rickettsial genetic conservation due to the ecologic separation and reduced selective pressure (Table 3)[112,172].

Although rickettsiae are maintained in nature within arthropod vectors, they frequently infect vertebrates, thereby allowing new lines of vectors to acquire infection from the infected hosts. The involvement of vertebrates is variable and in most cases humans are not essential in the rickettsial cycle. Only recently it was found that most arthropod *Rickettsia* are basal to the vertebrate *Rickettsia* and that the *Rickettsia* associated with leeches, protists and freshwater environments fall into two phylogenetic groups, distinct from the arthropod and vertebrate groups[113,153]. Ticks belonging to the family *Ixodidae* are the primary vector and reservoirs, and can also act as amplifiers of SFG *Rickettsiae* and *R. canadensis*. Although members of SFG are mainly associated with ticks, they can also associate with fleas and mites[148]. The specificity of association of these group of *Rickettsiae* with a particular tick species has been difficult to characterize, in particular regarding to how long a tick species has been associated with a rickettsial species and consequently if coevolution has followed[109]. However, different studies have already revealed that the relationship between SFG *Rickettsiae* and arthropods is relatively tick-specific. For example, some *Rickettsiae* such as *R. rickettsii,* may associate with ticks from different genera, whereas others such as *R. conorii,* appear to be associated with only one tick species. Between these extremes, there are few *Rickettsiae* which are associated with several tick species within the same genus, such as the association of *R. africae* and *R. slovaca* with various *Amblyomma* spp. and *Dermacentor* spp., respectively[173]. In contrast to SFG members, the primary vector of *R. prowazekii* transmission is the human body louse (*Pediculus humanus corporis*), although the presence in *Ambyomma* ticks is still undefined. Finally, fleas are the best known vectors of *R. typhi* and *R. felis* and mites are recognized vectors of *R. akari* and *O. tsutsugamushi*.

Mechanisms of infection

Rickettsioses are considered zoonotic diseases because they are transmissible from animal to humans and are considered vector-borne zoonoses due to the transmission by an assorted range of arthropods[107]. *Rickettsia* are stably maintained in nature in hematophagous arthropod vectors, but unstable when separated from host components. Nevertheless, as part of their life cycle, some species can also switch between arthropods and other secondary hosts, typically vertebrates (rodents, cattle, humans), while many other *Rickettsia* are found exclusively in arthropods with no known secondary host[172]. The host-preference patterns and modes of transmission are related to the infectious mechanism in the invertebrate host and, therefore, their geographic distribution is often determined by that of the infected arthropod[134].

With the exception of epidemic typhus, where humans play a crucial role in the life cycle of bacteria, usually rickettsiae do not infect humans during their natural cycles between arthropod and vertebrate hosts[109,174]. Their transmission to humans occurs accidentally through either direct inoculation into the skin of the host, from the feeding tick or mite's saliva during its blood meal, or contamination of broken skin and mucosal surfaces by feces of infected fleas or lice. Therefore, most rickettsial infections will not be transmitted from human to human or from human to non-human mammals.

It is widely believed that the infected insect feces are auto-inoculated into the skin of humans by scratching the skin irritated by the bite. Interestingly, extracellular *R. prowazekii* in louse feces and *R. typhi* in flea feces are stable and highly infectious, with the ability to survive within the feces for several weeks, if not longer[171]. Other forms of potential transmission of infectious *Rickettsiae* include the inoculation via rubbing the mucous membranes (e.g., conjunctiva) or via inhalation of aerosols. In fact, some tick-borne rickettsioses are transmitted by transfer of rickettsiae to the conjunctiva by fingers contaminated with infectious tick hemolymph or organs after crushing a tick that has been removed from a person or animal. Furthermore, aerosol transmission has been demonstrated experimentally to be very efficient, requiring 1000-fold fewer inhaled rickettsial organisms than anthrax spores[111].

Even though the knowledge about the life cycles of most tick-borne rickettsiae is still scarce, investigations over the last few years have suggested that the transmission of bacteria can be vertical, horizontal or both. In general, *Rickettsia* pathogenic species are transovarially transmitted (vertical transmission) to the next generation from infected tick female to offspring via the eggs, which allows many rickettsiae to be maintained in their arthropod hosts through generations (Figure 7)[175,176]. Another type of vertical transmission, the transstadial passage, in which the infection is maintained throughout different stages of the tick life cycle (from egg to larva to nymph to adult), is a necessary component for the vector competence of the ticks. When rickettsiae are transmitted efficiently both transstadially and transovarially in a tick species, it will serve as a reservoir of the bacteria and the distribution of the rickettsioses will be identical to that of its tick host[173]. For example, *R. slovaca* and *R. rickettsii* multiply in almost all organs and fluids of its tick host, particularly in the salivary glands and ovaries, which enables transmission of rickettsiae during feeding and transovarially, respectively[177,178]. Interestingly, recent studies of interspecies competition between different rickettsiae have shown that infection of a tick with one rickettsial species might alter the molecular-expression profiles of the oocytes, which interferes or block the infection by a second rickettsial species[176]. A last proposed mechanism of tick infection is horizontal transmission that is the acquisition of the bacteria by uninfected ticks feeding on infected animals[113,178]. As demonstrated with *R. conorii*, a particular type of horizontal transmission is co-feeding, which occurs when the uninfected vector gets infected through direct spread of bacteria from an infected tick during feeding at closely situated bite sites[179].

In contrast to most tick-borne rickettsiae, which are probably maintained in nature by all these mechanisms, flea- or louse-borne rickettsiae are not transmitted transovarially because they kill the vector that carries it. Thus, these bacteria have mammal reservoirs (humans for *R. prowazekii*, rodents for *R. typhi*, and cats for *R. felis*)[180] which directly spread bacteria by horizontal transmission. The fact that *R. prowazekii* is not motile can explain why it does not spread in its vector and why is the only *Rickettsia* sp. unable to be transmitted transovarially to its progeny in its vector[181].

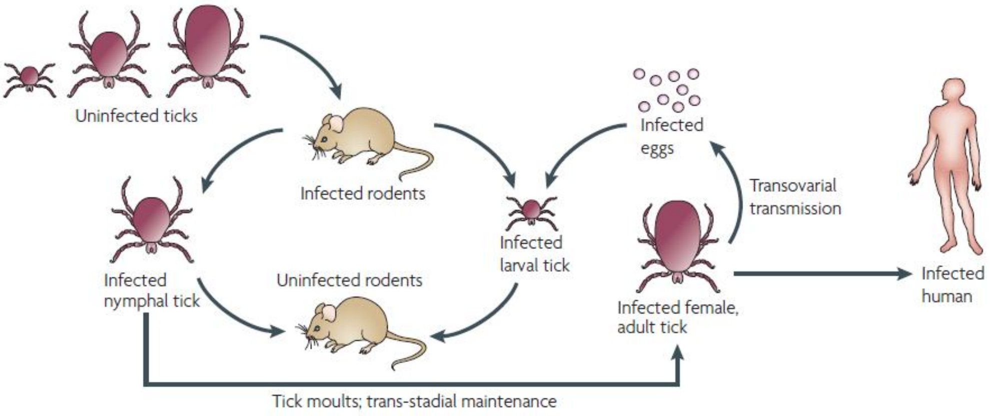


Figure 7. *SFG Rickettsia's life cycle*. The figure shows the transovarial and transstadial passage of SFG *Rickettsia* in the tick vector, as well as the horizontal transmission. Humans become incidental hosts after being bitten by an infected adult tick. Image from Walker and Ismail[108].

Virulence

Infectious diseases are major threats to human health worldwide and, in consequence, remarkable efforts have been dedicated into understanding numerous infectious agents and their pathogenic mechanisms. As most intracellular pathogens, rickettsiae developed highly specialized mechanisms to enter cells, cross cellular and biochemical barriers, and to overcome specific responses from the host organism. Many of these pathogenic steps are now firmly established and are the subject of the following sub-section.

Invasion

Because *Rickettsia* cannot replicate extracellularly within the mammalian host, these pathogens have developed sophisticated mechanisms to aid their adherence, replication and/or dissemination within nonphagocytic mammalian cells. Different studies have demonstrated that these three steps of bacterial infection are distinct events that are governed by specific protein-protein interactions at the pathogen-host cell interface. As with other pathogenic bacteria, the successful establishment of a rickettsial infection is highly dependent on the adherence to the host cell through an effective recognition and interaction of conserved rickettsial outer membrane-associated proteins with specific cellular receptors from the cells of the host, leading to eukaryotic downstream signaling and ultimately bacterial uptake[169].

A bioinformatics analysis of several sequenced rickettsial species, identified a family of at least 17 predicted outer surface proteins designated Sca (surface cell antigen) proteins that are predicted to encode proteins with homology to the autotransporter proteins of gram-negative bacteria[182]. Among these, the genes encoding rOmpA (Sca0), Sca1, Sca2, and rOmpB (Sca5), are conserved across the SFG[183–186], whereas the TG rickettsial species lack rOmpA and Sca2 gene appears to be fragmented in many of TG members[182]. Analysis of the amino acid sequence of these autotransporters highlighted a three domain organization for some members of this family: an N-terminal leader sequence that mediates transport across the cell membrane, a central passenger domain, and a C-terminal transporter sequence that is inserted as a β-barrel into the outer membrane to transport the passenger domain to the outer surface of the cell wall. Importantly, some of these proteins such as rOmpA and rOmpB, are translated first as preproteins and then processed to release the passenger domain from the β-barrel translocation domain, through a mechanism that is not clearly understood[183–185,187].

rOmpB, the most abundant rickettsial surface protein, has been shown to be highly conserved within either closely and distantly related rickettsial species, suggesting a similar function in the progression of different rickettsial diseases[188]. Chan and colleagues[186] have elucidated one major role for rOmpB during rickettsial infection using *E. coli* cells expressing a recombinant form of this protein. With this system, it was revealed that rOmpB from *R. conorii* and *R. japonica* are sufficient to trigger the adhesion and promote the bacteria uptake by nonphagocytic cells[186,189]. It was also found that the invasion process via rOmpB protein begins with the specific attachment to Ku70, a host cell protein that spans the membrane, leading to the recruitment of additional Ku70 molecules to the cell membrane, where further rOmpB binding occurs. The importance of rOmpB-Ku70 interaction was further confirmed by depleting Ku70 from mammalian cells, which were not permissive for bacterial invasion[186]. Despite this recognized role of rOmpB in rickettsial invasion process, studies involving the inhibition of rOmpB-Ku70 interaction disrupted *R. conorii* invasion of mammalian cells by approximately 50%, clearly suggesting that other surface proteins also contribute to this process[188,190]. So far, using similar approaches to examine the contribution of other proteins to early interactions of *Rickettsia*-target cells, rOmpA and Sca2 were also shown to mediate the attachment to host cells and to trigger rickettsial invasion, while Sca1 appears to be only capable of mediating the adhesion to mammalian cells[183–185]. Additional rickettsial adhesion proteins, encoded by the genes RC1281 and RC1282 in *R. conorii*, have also been proposed to be involved in bacterial adhesion and entry into the host cells[191].

Once attached to endothelial cells, signal transduction leads to actin rearrangement and to the actively internalization of *Rickettsia* pathogens in phagosomes, by a process defined as "induced phagocytosis". An electron microscopy analysis of rickettsial entry suggested that rickettsial invasion of normally non-phagocytic cells is morphologically and mechanistically related to a "zipper-like" invasion strategy. Contrary to the trigger mechanism, the alternative pathway utilized by other invasive bacteria, in the zipper mechanism specific bacterial receptor-ligand interactions (e.g., rOmpB-Ku70) induce focal actin recruitment and progressive apposition of the plasma membrane over the bacterium[170,192]. An investigation into the molecular details responsible for the remodeling of the actin cytoskeleton during bacterial entry into host cells, revealed that *R. conorii* recruits components of the Arp2/3 complex to the site of entry foci[169]. Different approaches used to disrupt signaling pathways that directly or indirectly activate the Arp2/3 complex revealed that *R. conorii* utilizes pathways involving Cdc42, phosphoinositide 3-kinase, c-Src, cortactin and other protein tyrosine kinase activities to enter non-phagocytic cells[169]. Ubiquitination of Ku70 via c-Cbl and the involvement of clathrin and caveolin 2 implicate the host endocytic machinery in the invasion pathway (Figure 8)[190].

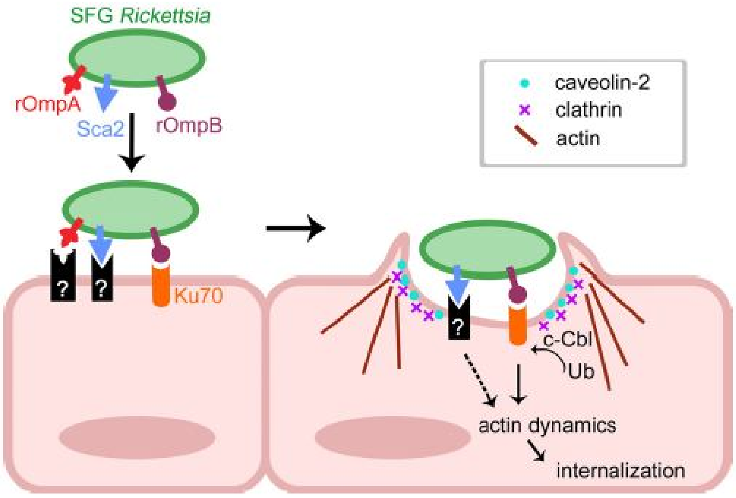


Figure 8. *Adhesion and invasion mechanisms of SFG rickettsiae.* SFG rickettsiae adhere to host cell membrane through interaction of rOmpB with its membrane embedded receptor, Ku70. Besides this interaction it has also been proved that Sca2 is involved in this process, being its receptor still unknown (black box). There are some evidences that despite acting in different cell receptors, both rOmpB and Sca2 trigger a signal cascade ultimately converging to localized recruitment of actin filaments and endocytic machinery (clathrin, caveolin-2 and c-Cbl) to entry into the host cell. Image from Cardwell and colleagues[185].

Phagosomal Escape

Following internalization of rickettsiae, bacteria must escape from phagosome into the cytoplasm prior to phagolysosomal fusion, thereby avoiding exposure to lysosomal enzymes. This mechanism has been suggested to be dependent on proteins with membranolytic activity that can digest the host cell phagosomal membrane, including phospholipase A2 (PLA2), hemolysin C (Tlyc) and phospholipase D (PLD)[193]. PLD from *R. conorii* and *R. prowazekii* were the first phospholipases identified within rickettsial genomes, and *in vitro* studies have revealed that this gene is functional[194]. Later studies also confirmed that this gene is conserved in all species of the *Rickettsia* sequenced up to now, and that PLD is likely the major effector of rickettsial phagosomal escape[193,195]. Although the involvement for a PLA2 in the entry vesicle lysis has for long been anticipated for *R. rickettsii* and then extended to both *R. conorii* and *R. prowazekii*[196], it was not until recently that genes encoding PLA2 homologues were found in the *R. typhi, R. prowazekii, R. massiliae* and *R. bellii* genomes[197]. This study also suggested that PLA2 is secreted into the host cytoplasm and provided additional support for the notion that PLA2 is a bona fide enzyme with functional phospholipase activity in *R. typhi*[197]. In addition to phospholipases, the membrane-disrupting TlyC from *R. rickettsii* and the homologue from *R. prowazekii* were also shown to have hemolytic activity on normally nonhemolytic bacteria[193,198]. Nevertheless, a direct role for TlyC in rickettsial phagosome escape has not yet been demonstrated.

Actin-based motility

Once in the cytoplasm, rickettsiae acquire their survival nutrients from the host cell (ATP, amino acids, nucleotides), allowing them to grow and replicate inside the host. Investigations on the intracellular life-style of rickettsiae noted that like several other microbial pathogens, rickettsiae explore the host-cell actin cytoskeleton to enter and to disseminate within cells, thus avoiding the host immune response[169,199]. The filaments of actin push the *Rickettsia* to the surface of the host cell, where the host cell membrane is deformed outward and invaginates into the adjacent cell[200]. Historically, this intracellular spreading mechanism was described as the major feature allowing differentiation of SFG from TG rickettsiae. Members of the TG are not motile within the cells and the infection of adjacent cells only happens when the bacterial load increases (5-8 times greater than that observed for SFG) and induces the lysis of host cells. Conversely, most SFG rickettsiae exploit the host cell actin cytoskeleton to promote intracellular mobility via active propulsion by means of directionally polymerized actin, culminating in extensive membrane damage and eventual cell death[117,181]. The molecular mechanisms of actin polymerization primarily involve the expression of a surface WASP-like protein, RickA, which activates the Arp2/3 complex, an actin nucleator responsible for initiating the polymerization of new actin filaments in the cytoplasm at one pole of the bacteria and the induction of the formation of a network of long and unbranched actin tails in *Rickettsia* spp.[200]. Consistent with the importance of RickA in SFG rickettsial actin-based mobility, *RickA* gene was found to be absent in *R. prowazekii* genome, thereby clarifying the absence of motility on members of TG[161]. In addition to RickA, Sca2 protein was also recently found to be also implicated in actin assembly during actin-based motility. In this study, a random transposon mutagenesis of *R. rickettsii* disrupted the Sca2 gene by a transposon insertion causing a small plaque phenotype[201]. Importantly, a strong suggestion that Sca2 is a virulence determinant for SFG rickettsiae arose with the observation that, in a guinea pig model, the Sca2 mutant does not induce fever as does the congenic wild-type strain[201].

Host Injury and Immune Response

For the majority of the members from the SFG, infection of vertebrates starts with the attachment of the ticks and mites which then prepare to imbibe a blood meal that takes place over a period of 3 to 14 days after biting. *Rickettsia* are then transmitted by infected ticks to humans through their saliva during blood feeding. While the tick salivates into the wound, a number of secreted proteins including enzymes, vasodilators, antihemostatic and immunomodulatory substances are thought to influence innate and adaptive immune responses in the skin. During this stage, the low immune activity at the bite site prevents the host from rejecting the ticks and enhances the transmission potential of rickettsiae harbored either by vector or host[171,202]. Soon after the tick bite, localized inoculation of rickettsiae promotes the tissue damage, giving rise to a necrotic lesion known as eschar[203].

While the primary targets of *Rickettsia* are endothelial cells [with the exception for *R. akari* (the agent of Rickettsialpox) which primarily infects macrophages], as bacteria spread they are able to infect and proliferate within any nucleated cell. Therefore, secondary targets include dermal cells such as fibroblasts, macrophages, dermal dendritic cells, lymphatic endothelium and, for *R. rickettsii* in RMSF perivascular smooth muscle cells[202]. Upon entry into host skin, *Rickettsia* is rapidly spread all over the body host through dissemination via bloodstream and lymphatic circulation, followed by the infection and damage to endothelial cells of the small capillary beds of many organs including the lungs, brain, liver, heart and kidney[171].

The most prominent pathophysiological effect of rickettsial infection is characterized by an increased microvascular permeability due to the disruption of adherens junctions, consequent development of gaps, formation of stress fibers, and conversion of the shape of endothelial cells from polygons to large spindles[204–206]. The subsequent increased fluid leakage into the interstitial space and further infiltration of perivascular mononuclear cells often results in a characteristic dermal rash[110,207]. In addition, endothelial dysfunction and activation is followed by acute phase responses characterized by generalized vascular inflammation, edema, increased leukocyte–endothelium interactions and release of powerful vasoactive mediators that promote coagulation and pro-inflammatory cytokines, all features collectively termed as "rickettsial vasculitis" [108,203,208]. As the disease progresses, organ and tissue damage due to loss of blood into tissue spaces can become a life threatening situation, especially in organs where there are no lymphatic vessels to remove interstitial fluid, such as brain and lungs. Severe complications in untreated cases with widespread vasculitis can include encephalitis, noncardiogenic pulmonary edema, interstitial pneumonia, hypovolemia, hypotensive shock, and acute renal failure, responsible for the high morbidity and mortality associated with rickettsioses[110,137,171].

The mechanisms underlying the host defense are not yet completely understood, although both humoral and cell mediated immunity are thought to play a crucial role in recovery from infection. Most of our understanding of the immune response against *Rickettsia* is derived from *in vitro* studies, as well as from murine models of SFG and TG rickettsioses which have identified novel mechanisms of immunity, including cytokine-mediated activation of endothelial cell bactericidal control of intracellular infection and the role of autophagy in rickettsial killing[110].

Early in rickettsial infection process, innate immune responses are thought to limit the growth and spread of rickettsiae through activation of natural killer cell activity in association with the production of IFN-γ[209]. This particular type of interferon is responsible for inducing antiviral function and for activating macrophages and dendritic cells in the presence of IL-12 and IL-18 to support innate immune responses to microorganisms. Further migration to the foci of infection of perivascular CD4 and CD8 T cells, macrophages and dendritic cells, is presumed to promote clearance of rickettsiae through activation of intracellular bactericidal mechanisms on endothelial cells, most likely by secreting pro-inflammatory cytokines and chemokines. Human endothelial cells activated by IFN-γ, TNF-α, RANTES and IL-1b, kill intracellular rickettsiae via nitric oxide and hydrogen peroxide production through nitric oxide synthesis-dependent and indoleamine 2,3-dioxygenase-dependent mechanisms [202,210]. On the other hand human macrophages, a minor target of rickettsiae, kill intracellular bacteria after activation by IFN-γ, TNF-α, and IL-1b via production of hydrogen peroxide and tryptophan starvation of rickettsiae associated with degradation of tryptophan by indoleamine-2,3-dioxygenase. The pathogenic mechanism of oxidative stress associated with *R. rickettsii* injury has been proven to cause host cell membranes lipid peroxidation and was shown to be associated with depletion of host components such as glutathione and increased levels of catalase. These phenomena increases the concentration of hydrogen peroxide and leads to a striking reduction in enzymes such as glucose-6-phosphate dehydrogenase, glutathione peroxidase, and catalase that are host defenses against ROS-induced damage[171,211].

A secondary effector component of the acquired immune response against *Rickettsia* is the generation of specific cytotoxic CD8+ T cells that induce apoptosis in infected target cells via pathways involving perforin and/or granzymes. Cytotoxic activity of CD8 T-lymphocytes is crucial to the clearance of rickettsial infection[212]. Rickettsial manipulation of its host cell also includes activation of NF-kB as one major strategy employed by rickettsiae to survive and replicate within the endothelium. Activation of NF-kB inhibits endothelial cell apoptosis by preventing apical activation of caspases-8, -9 and -3 and also mediates the production of proinflammatory cytokines and chemokines, such as IL-8, IL-6, IL-1α[213–215]. However, later in the infectious process, when both innate and active immunity are fully activated, the anti-apoptotic effect of NF-kB is likely overridden by the cytotoxic CD8+ T-induced apoptosis of infected endothelial cells[215].

The humoral response may play an important role in protection against infection and antibodies against rickettsial OmpA and OmpB, but not rickettsial lipopolysaccharide, are indeed protective against re-infection[216,217]. However, antibodies towards these proteins do not appear until the control and recovery from the disease has occurred. Thus, antibodies may be more important in preventing re-infection and in vaccine-induced immunity than in clearance of primary infection[218].

Diagnostics and Therapeutics

Although important progress has been made on the fields of molecular biology, cellular biology, and immunology and pathogenesis of *Rickettsia*, diagnosis of rickettsial diseases is still difficult and is usually retrospective. Rapid diagnostic methods are still required for diagnosis of rickettsial diseases and in response to their potential use in bioterrorism.

Rickettsioses can present an array of clinical signs and symptoms which generally manifest 2–14 days following *Rickettsia* inoculation. These diseases vary in severity from self-limited mild infections to fulminating life-threatening diseases, but are generally characterized by acute onset of high fever, which may last up to 2 weeks as in the case of TG rickettsioses. Other symptoms may include severe headache, prominent neck muscle myalgia, malaise, nausea/vomiting, or neurologic signs[203]. The characteristic macular or maculopapular rashes appear 3-5 days following onset of the disease in most patients (~90%) infected with RMSF or epidemic typhus. However, in other less severe spotted fevers, such as African tick bite fever and *R. parkeri* infection, rash may be less frequent (10-15%). Conversely, focal skin necrosis with a dark scab (eschar) at the site of tick feeding is a common feature of MSF, African tick bite fever, North Asian tick typhus, Queensland tick typhus, Japanese spotted fever, Flinders Island spotted fever, Rickettsialpox, tick-borne lymphadenopathy, and the recently described infections in the US caused by *R. parkeri* and by a novel strain 364 D, but is rare in RMSF[203]. While these symptoms aid proper diagnosis of the infectious agent, the disease often manifests itself as nondescript fever and flu-like symptoms, leading to misdiagnosis and inappropriate treatment. Misdiagnosis of *Rickettsia* infection is associated with severe disease consequences, including interstitial pneumonia, neurological pathology, acute renal failure, pulmonary edema, and other multiorgan manifestations. Untreated MSF and RMSF can result in mortality rates estimated to be as high as 23%, but appropriate treatment drastically decreases the risk[109].

The early diagnosis of rickettsial diseases is based mostly on clinical suspicion since no reliable diagnostic test is available on the early phase of the illness. When the disease is clinically suspected, biological diagnosis can be obtained using serology, cell culture and/or molecular tools[109]. Until recently, the diagnosis of rickettsioses was confirmed almost exclusively by serological tests. Of the serological tests, the indirect micro-immunofluorescence assay has been the most sensitive and specific, but usually IgM and IgG antibodies reactive with *Rickettsia* are undetectable during the first week of illness[219]. In addition, there is an extensive antigenic cross-reaction among SFG and TG rickettsiae, making immunofluorescence assays a less helpful tool to distinguish between the species[220]. Other serologic tools are the Weil-Felix test, complement fixation, micro-agglutination test, latex agglutination, ELISA, and Western immunoblot assays[221].

Among the molecular tools, real-time quantitative PCR (qPCR) has been claimed the most rapid and sensitive, while reducing the costs and the time of diagnosis[222,223]. Several genes are commonly used for detection of rickettsial DNA such as the *Rickettsia* genus specific 17-kDa antigen gene, the 16S rRNA gene, the citrate synthase gene (gltA), and the outer membrane proteins rOmpB and rOmpA.

Once diagnosed, bacteriostatic antibiotics from the tetracycline class (specifically doxycycline) are the most widely used and normally effective in treating rickettsioses. Depending on the scenario, chloramphenicol, azithromycin, fluoroquinolones, and rifampin may be used as alternatives to doxycycline. The treatment with the proper antibiotic should be initiated immediately after a suspicion of rickettsial infection, and must be continued for at least 3 days after fever diminishes and until there is a clear evidence of clinical improvement[224].

Although there is an increasing worldwide concern with human infections caused by the genus *Rickettsia*, relatively little is known about the factors that are required to elicit a protective immune response. The need for a reliable protective vaccine to prevent rickettsial infections is well recognized and a number of vaccine candidates have been tested with varying degrees of success. In the past, prospect for developing effective killed whole-cell and live attenuated vaccines against rickettsial diseases have culminated in numerous failures and limited success in preventing or ameliorating the disease[110]. Even though, recent studies have shown that after recovery from *R. rickettsii* and *R. conorii* infections, patients and experimental animals develop solid immunity that prevents reinfection, thus indicating that stimulation of protective immunity is entirely feasible. Current challenges are focused on the identification of rickettsial proteins that stimulate the components of the immune system that confer protection, both cellular and humoral[225]. The development and validation of an effective vaccine against any of the rickettsioses that could provide adequate prophylaxis would definitely reduce the impact of these diseases worldwide.

1.3. Research objectives

The recently observed increase in the reported incidence of rickettsial infections worldwide is a cause for renewed concern, as it is the potential fatal outcome associated with these infections. The life-threatening character of many *Rickettsia* spp. results from their highly virulent properties and unique biological characteristics including the associated high morbidity and mortality, environmental stability, aerosol transmission, persistence in infected hosts and low infectious dose. Furthermore, the emerging character of rickettsioses together with the difficulties of diagnostics, the lack of reliable protective vaccine and their potential use as bioterrorism weapons, strengthens the importance of identifying new protein factors for the potential development of innovative tools to prevent, diagnose and treat these infections diseases[110,111].

In this perspective, as part of the core genome of all *Rickettsia* sequenced so far, we have identified a putative gene encoding a membrane-embedded aspartic protease with a retroviral-type AP signature, for which no function has been assigned. Using *R. conorii*'s AP (RC1339) as our working-model, hereby named APRc for Aspartic Protease from *Rickettsia conorii,* the goal of the present study was to provide a comprehensive biochemical and enzymatic characterization of this novel retroviral-type AP and also give further insights on the putative functional role of this enzyme. The prospects opened with this work will pave the way, more broadly, to further research on how rickettsial AP might contribute for the pathogenicity of these parasites, with special emphasis on its potential use as a target for therapeutic intervention in rickettsioses. Moreover, although evidences do exist for the occurrence of APs in bacteria, this is the first report on this class of enzymes in gram-negative intracellular species like *Rickettsia*, thereby giving a valuable contribution for the discussion on the evolution of APs.

Accordingly, in Chapter I is presented and discussed the recombinant expression of the soluble catalytic domain of APRc, using *E. coli* as the heterologous production system. A detailed enzymatic and biochemical characterization of this protease is provided with respect to the auto-processing activity, dimerization and enzymatic properties, and the similarities and differences with other well studied aspartic proteases are evaluated.

In order to determine the specificity profile of APRc, we applied the innovative technique Proteomics Identification of protease Cleavage Sites (PICS) in collaboration with Dr. Christopher Overall from the University of British Columbia, Canada. A comparative analysis with the specificity of other APs was also carried out (Chapter II).

Finally, Chapter III provides initial evidences for the potential biological relevance of this protease in rickettsial life-cycle. First, we addressed the type of association, location and topology of full-length APRc with *E. coli* membranes. Next, we focused on the expression and location analysis of native APRc in different rickettsial strains and on the evaluation of Sca proteins as the putative substrates for this protease. Because manipulation of pathogenic *Rickettsiae* from host cells requires appropriate biosafety level 3 facilities, part of these studies were conducted in collaboration with Dr. Juan Martinez, from the Louisiana State University, USA.

Chapter II. Biochemical and enzymatic characterization of APRc

Chapter II. Biochemical and enzymatic characterization of APRc

2.1. Introduction

Bacterial pathogenicity generally results from a combination of factors where different bacterial components and strategies contribute to virulence[226]. Among these components, a diverse array of proteolytic enzymes (mainly localized to the bacterial surface or secreted) have been recognized as virulence factors in several pathogenic bacteria by playing critical functions related to colonization and evasion of host immune defenses, acquisition of nutrients for growth and proliferation, facilitation of dissemination or tissue damage during infection[226–228]. The relevance of proteolytic events for bacterial pathogenicity and the progressive increase in antibiotic resistance among pathogenic bacteria contribute to position proteases as potential candidate targets for the development of alternative antibacterial strategies[228].

In line with what has been described for other obligate intracellular bacteria, rickettsial species have highly conserved and reduced genome sizes, which derive from reduction of originally larger genomes accompanying the adaptation to strict intracellular lifestyles[136,159,162,163]. Although significant progress has been made concerning both genotyping and epidemiology of rickettsiae, the genetic intractability of these bacteria has severely limited molecular dissection of virulence factors associated with their intracellular parasitism and pathogenic mechanisms. However, the availability of complete genome sequences of a vast number of species works as an invaluable tool to unmask hidden proteases, thereby providing new sets of potential targets. On this subject, though, with the exception for a few examples belonging to the secretory pathway (e.g., type I[229], type II[230] and type IV[231,232] signal peptidases) and a TG-specific prolyl oligopeptidase protein[233], no other *Rickettsia* proteases have been identified and characterized. A major challenge for the future rests then in the demonstration and characterization of enzymatic activity, properties and function of a myriad of *in silico* predicted proteases.

Herein, we describe the identification of a gene coding for a putative membrane-embedded aspartic protease (AP) of the retropepsin-type, conserved in all 55 sequenced *Rickettsia* genomes. The retropepsins were first identified with the discovery of the HIV-1 PR in the late 1980's[234] and the recognition of its essential role in the maturation of HIV-1. As previously mentioned, these proteases require homodimerization of two monomeric units in order to form a functional enzyme, structurally related to the pepsin family (A1) of bilobal APs[27,36,41]. Strikingly, the presence of retropepsins in prokaryotes has long been a matter of debate but never unequivocally demonstrated. The work presented in this chapter provides a detailed description on the identification and characterization of the retropepsin homologue from *Rickettsia conorii* (RC1339/APRc) and demonstrates that this protease is active and shares several enzymatic properties with other members of this family of APs (e.g., autolytic activity, optimum pH, and sensitivity to specific HIV-1 PR inhibitors).

2.2. Materials and Methods

Materials

Oligonucleotide primers were purchased from Integrated DNA Technologies, Leuven, Belgium. Synthetic genes encoding the full-length RC1339 and the predicted soluble catalytic domain, the fluorogenic peptides PepRick14 (MCA-Lys-Ala-Leu-Ile-Pro-Ser-Tyr-Lys-Trp-Ser-Lys-DNP), PepRick15 (MCA-Lys-His-Arg-Val-Met-Ser-Ala-Leu-Ile-Lys-DNP) and the rabbit polyclonal antibody raised towards the sequence Cys-Tyr-Thr-Arg-Thr-Tyr-Leu-Thr-Ala-Asn-Gly-Glu-Asn-Lys-Ala (anti-APRc) were produced by GenScript (Piscataway, NJ, USA). N-terminal amino acid sequence analyses were performed in the Analytical Services Unit - Protein Sequencing Service, ITQB (Oeiras, Portugal). Circular dichroism analyses were performed at Applied Photophysics on a Chirascan plus Auto-CD.

Bioinformatics analysis

Gene and protein sequences for *R. conorii* str. Malish 7 RC1339 were obtained from the genome sequence at NCBI (NC_003103) (AAL03877). Amino acid sequence alignment and the degree of identity between RC1339/APRc homologues from *Rickettsia* (genus) (TaxID 780) were obtained with ClustalW[235], by comparing the 55 sequences deposited in NCBI database. The protein family, domain, and functional sites were searched using the InterProScan program[236]. Topology structure was predicted with HMMTOP2 algorithm[237]. A structure-based alignment of RC1339/APRc soluble catalytic domain with HIV-1 (PDB 3hvp), equine infectious anemia virus (EIAV) (PDB 2fmb) and XMRV (PDB 3nr6) retropepsins and with Ddi1 putative protease domain (PDB 2i1a) was performed with PROMALS3D[238].

DNA constructs

The sequence encoding the full-length of RC1339/APRc (construct coding amino acids 1-231) was chemically synthetized with OptimumGene™ codon optimization technology to *E. coli* codon usage and cloned into pUC57 vector. The sequence was then amplified to include restriction sites for NcoI and NotI at 5'- and 3'-ends, respectively, using the forward primer 5'-CCATGGGAATGAACAAAAAACTGATCAAACTG-3' and the reverse primer 5'-CTCGAGATAATTCAGAATCAGCAGATCTTT-3'; the resulting PCR product was cloned into pGEM-T

Easy plasmid (Promega). After digestion with NcoI and NotI, the insert was subcloned into pET28-a expression vector (Invitrogen) in frame with a C-terminal His-tag (pET-APRc$_{1-231}$-His). In order to generate the untagged construct, an insertion mutagenesis was performed to include the TGA stop codon at the end of the full-length sequence using the Quick Change site-directed mutagenesis kit (Stratagene) and the primers 5'-ATTCTGAATTATTGACTCGAGCACCAC-3' (forward) and 5'-GTGGTGCTCGAGTCAATAATTCAGAAT-3' (reverse) (pET-APRc$_{1-231}$) (Supplemental Figure 1A).

The coding sequence for the predicted soluble catalytic domain of RC1339/APRc (construct coding amino acids 87-231) was chemically synthetized with OptimumGene™ codon optimization technology to *E. coli* codon usage and cloned into pUC57 vector. The sequence flanked by restriction sites for BamHI (5')/EcoRI (3') was then inserted in frame to the C terminus of GST (glutathione S-transferase) in pGEX-4T2 expression vector (Amersham) using the same pair of restriction enzymes (pGST-APRc$_{87-231}$ (Supplemental Figure 1D)).

For generating the expression constructs bearing the sequence encoding the intermediate activation form rAPRc$_{99-231}$-His and the final activated form rAPRc$_{105-231}$-His, both sequences were firstly amplified using the construct pET-APRc$_{1-231}$-His as the template and the forward primer containing a NdeI restriction site (5'-CATATGAGCGCCCTGATCCCGTCT-3' for pET-APRc$_{99-231}$-His and 5'-CATATGTATAAATGGAGTACCGAAGTT-3' for pET-APRc$_{105-231}$-His) and the same reverse primer used for amplification of pET-APRc$_{1-231}$-His (5'-CTCGAGATAATTCAGAATCAGCAGATCTTT-3'), and cloned into pGEM-T Easy (Promega). The inserts were then digested with NdeI/NotI and subcloned into pET23a expression vector (Invitrogen) in frame with a C-terminal His-tag (pET-APRc$_{99-231}$-His (Supplemental Figure 1F) and pET-APRc$_{105-231}$-His (Supplemental Figure 1J)).

The active site mutants of APRc constructs rGST-APRc$_{87-231}$ [pGST-APRc(D140A)$_{87-231}$ (Supplemental Figure 1B)] and rAPRc$_{99-231}$-His [pET-APRc(D140A)$_{99-231}$-His (Supplemental Figure 1G)], were generated by replacing the putative active site aspartic acid residue by alanine (D140A) using the Quick Change site-directed mutagenesis kit (Stratagene) and the primers 5'-AAAATCAAATTCATGGTGAATACCGGCGCCTCTGATATTGCA-3' (forward) and 5'-TGCAATATCAGAGGCGCCGGTATTCACCATGAATTTGATTTT-3' (reverse) (mutation underlined).

Two mutants of the third auto-catalytic cleavage site identified upon activation of rGST-APRc$_{87-231}$ were also produced using the construct pET-APRc$_{99-231}$-His as template and the Quick Change site-directed mutagenesis kit (Stratagene) for the substitution of the Ser104 and Tyr105 by a proline [constructs: pET-APRc(S104P)$_{99-231}$-His (Supplemental Figure 1H) and pET-APRc(Y105P)$_{99-231}$-His (Supplemental Figure 1I), respectively]. For the mutation of Ser104 the following pair of primers were used 5'-

CATATGAGCGCCCTGATCCCGCCTTATAAATGGAGTACCGAAG-3’ (forward) and 5’-CTTCGGTACTCCATTTATAAGGCGGGATCAGGGCGCTCATATG-3’ (reverse), whereas for Tyr105 the primers were 5’-GAGCGCCCTGATCCCGTCTCCTAAATGGAGTACCGAAGTTG-3’ (forward) and 5’-CAACTTCGGTACTCCATTTAGGAGACGGGATCAGGGCGCTC-3’ (reverse) (mutation underlined).

All positive clones were selected by restriction analysis and confirmed by DNA sequencing.

Expression and purification of the soluble forms of APRc

rGST-APRc$_{87\text{-}231}$ and the corresponding active site mutant protein were expressed by standard procedures. Briefly, *E. coli* BL21 Star (DE3) cells transformed with each plasmid construct, pGST-APRc$_{87\text{-}231}$ and pGST-APRc(D140A)$_{87\text{-}231}$, were grown at 37 ˚C until an OD$_{600nm}$ of 0.7. Protein expression was then induced with 0.1 mM IPTG for 3 h, after which cells were harvested by centrifugation at 9000*g* for 20 min at 4 ˚C, and resuspended in PBS buffer. Lysozyme (100 µg/mL) was added and the harvested cells were frozen at -20 ˚C. After freezing and thawing, bacterial cell lysates were incubated with DNase (1 µg/mL) and MgCl$_2$ (5 mM) for 1 h at 4 ˚C. The total cell lysate was then centrifuged at 27216*g* for 20 min at 4 ˚C and the resulting supernatant filtered (0.2 µm) before loading onto a GSTrap HP 5 mL column (GE Healthcare Life Sciences) previously equilibrated in PBS buffer. After extensive washing, the protein of interest was eluted in 50 mM Tris-HCl pH 8 with 10 mM glutathione and immediately loaded onto a Superdex 200 HiLoad 26/60 (GE Healthcare Life Sciences) equilibrated in PBS buffer for further purification and glutathione removal.

Expression of *E. coli* BL21 Star (DE3) cells transformed with pET-APRc$_{99\text{-}231}$-His plasmid as well as isolation of total soluble protein were performed under the same conditions as described for rGST-APRc$_{87\text{-}231}$, except that in this case the cell pellet was resuspended in 20 mM phosphate buffer pH 7.5, 500 mM NaCl and 10 mM imidazole. The resultant supernatant was then loaded onto a HisTrap HP 5 mL column (GE Healthcare Life Sciences) pre-equilibrated in the same buffer. Protein elution was performed by a three-step gradient of imidazole (50 mM, 100 mM and 500 mM) and fractions containing the protein of interest (100 mM Imidazole gradient step) were pooled and buffer exchanged into 20 mM phosphate buffer pH 7.5 by an overnight dialysis step. Dialyzed protein was further purified by cation-exchange chromatography with a MonoS 5/50 column (GE Healthcare Life Sciences) equilibrated in the same buffer and elution was carried out by a linear gradient of NaCl (0-1 M).

Autoprocessing activity of APRc in soluble extracts of *E. coli*

E. coli BL21 Star (DE3) cells transformed with pGST-APRc$_{87-231}$ (or pGST-APRc(D140A)$_{87-231}$) were grown at 37 ˚C until an OD$_{600nm}$ of 0.7. At this point, 1 mL of cell culture was subjected to induction of protein expression with 0.1 mM IPTG for 3 h. Cells were then harvested by centrifugation at 10000*g* for 20 min at 4 ˚C, and resuspended in 200 µL of the protein extraction reagent BugBuster (Merck Millipore). After 20 min, extracts were clarified by centrifugation and analyzed by Western blot with anti-APRc antibody.

In vitro auto-processing studies

The autoproteolytic activity of rGST-APRc$_{87-231}$ was primarily assessed over a pH range of 3.0 to 7.0. APRc purified samples were diluted 1:1 with 0.1 M sodium citrate buffer pH 3, 0.1 M sodium acetate buffer pH 4, 0.1 M sodium acetate buffer pH 5, 0.1 M sodium acetate buffer pH 5.5, 0.1 M sodium acetate buffer pH 6 and 0.1 M Tris-HCl pH 7, and incubated for approximately 24 h for SDS-PAGE analysis (silver staining).

Time-course studies of APRc activation were undertaken with two recombinant forms of the soluble catalytic domain of APRc (rGST-APRc$_{87-231}$ and rAPRc$_{99-231}$-His). Purified samples of APRc were first diluted to 0.1 mg/mL with PBS and then diluted 1:1 with 0.1 M sodium acetate buffer pH 6. Diluted samples were incubated up to 48 h at 37 ˚C and aliquots were taken every 12 h for SDS-PAGE analysis and proteolytic activity assays. To evaluate the effect of inhibitors on APRc auto-activation processing, a time-course analysis was carried out in the presence of 20 µM pepstatin, 1 mM indinavir or 5 mM EDTA and protein samples analyzed by SDS-PAGE.

Analytical size-exclusion chromatography

Precursor rAPRc$_{99-231}$-His and activated APRc$_{110-231}$-His forms were analyzed under nondenaturing conditions by analytical size-exclusion chromatography (SEC) on a Superdex 200 5/150 GL (GE Healthcare Life Sciences) column connected to a Prominence HPLC system (Shimadzu Corporation, Tokyo, Japan). The column was equilibrated in 20 mM phosphate buffer pH 7.5 containing 150 mM NaCl, and calibrated with Gel Filtration LMW and HMW calibration kits (GE Healthcare Life Sciences), according to the manufacturer's instructions. The molecular mass markers used for calibration were conalbumin (75 kDa), ovalbumin (43 kDa), carbonic anhydrase (29 kDa), and ribonuclease A (13.7 kDa).

Dimerization studies

Cross-linking reactions with disuccinimidyl suberate (DSS) (Pierce) were performed in 20 mM phosphate buffer pH 7.5 containing 150 mM NaCl. A solution of 0.2 mg/mL of purified APRc (rAPRc$_{99\text{-}231}$-His and activated product APRc$_{105\text{-}231}$-His) was treated with a 50-fold molar excess of DSS in a total volume of 50 µl and allowed to react for 30 min at room temperature. For glutaraldehyde treatment, a solution of 0.5 mg/mL of purified rAPRc$_{99\text{-}231}$-His was treated with 5 µl of 1.15% freshly prepared solution of glutaraldehyde for 4 min at 37 ˚C, in a total volume of 50 µl, under similar buffer conditions. To terminate the reactions, 5 µl of the quenching buffer 1 M Tris-HCl pH 8.0 were added. Crosslinked proteins were separated by SDS-PAGE and analyzed by Western blot with anti-APRc antibody and also analyzed by analytical SEC as previously described.

Enzyme activity assays

Activity screening with Insulin β-chain

Evaluation of proteolytic activity during auto-processing time-course analysis was performed towards oxidized insulin β chain. Substrate (1 mg/mL) was incubated with purified recombinant APRc enzyme (corresponding to different times points of activation of rGST-APRc$_{87\text{-}231)}$): substrate mass ratio of 1:15) in 0.1 mM sodium acetate buffer pH 6.0. After an overnight incubation at 37 ˚C the reaction mixture was centrifuged at 20000*g* during 6 min and the digestion fragments were separated by RP-HPLC (reversed-phase high performance liquid chromatography) on a C18 column (KROMASIL 100 C18 250, 4.6 mm), using a Prominence system (Shimadzu Corporation, Tokyo, Japan). Elution was carried out with a linear gradient (0–80%) of acetonitrile in 0.1% v/v trifluoroacetic acid for 30 min at a flow rate of 1 mL/min. Absorbance was monitored at 220 nm.

Enzymatic characterization by fluorescence activity assays

The effect of pH on activity and inhibitory profile of activated APRc (APRc$_{105\text{-}231}$-His) was determined by fluorescence assays in 96-well plates in in a Gemini™ EM Fluorescence Microplate Reader, using the fluorogenic substrate PepRick14 ([MCA]-Lys-Ala-Leu-Ile-Pro-Ser-Tyr-Lys-Trp-Ser-Lys-[DNP]) (final concentration of 2.5 µM). For determination of the pH profile, APRc$_{105\text{-}231}$-His was assayed for activity at 37 ˚C in buffers ranging between pH 4 and 9 (50 mM sodium acetate pH 4.0, 5.0, 5.5 and 6.0; 50 mM Tris-HCl pH 7.0, 8.0 and 9.0) containing 100 mM NaCl. To test the effect of classical inhibitors, the protease was pre-incubated in the presence of each inhibitor, 20 µM pepstatin, 5 mM EDTA, 1 mM $ZnCl_2$, 1 mM Pefabloc, or 10 µM E-64, for 10 min at room temperature in 50 mM sodium acetate pH 6.0 containing 100 mM NaCl before determination of proteolytic activity. The effect of the HIV-1 PR inhibitors on APRc proteolytic activity was also evaluated. The following reagents were obtained through the NIH AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH: indinavir sulfate, nelfinavir, ritonavir, saquinavir, amprenavir, atazanavir sulfate, darunavir, and lopinavir. Each inhibitor was again incubated with APRc for 10 min at room temperature in 50 mM sodium acetate pH 6.0 containing 100 mM NaCl and 5% DMSO, except for indinavir and darunavir that were assayed without DMSO. Indinavir, nelfinavir, ritonavir, saquinavir, amprenavir, atazanavir and lopinavir were tested in the range of 0.25 – 1 mM and the inhibitor darunavir in the range of 2.5 – 10 µM (final concentration).

The enzymatic activity of active APRc was also tested towards five different fluorogenic peptides: PepRick15 ([MCA]-Lys-Tyr-His-Arg-Val-Met-Ser-Ala-Leu-Ile-Lys-[DNP]), typical AP substrate ([MCA]-Lys-Lys-Pro-Ala-Glu-Phe-Phe-Ala-Leu-Lys-[DNP]), BACE1 substrate ([MCA]-Leu-Ser-Glu-Val-Asn-Leu-Asp-Ala-Gly-Phe-Lys-[DNP]), HIV-1 PR substrate (Arg-Glu-[EDANS]-Ser-Glu-Asn-Tyr-Pro-Ile-Val-Gln-Lys-[DABCYL]-Arg) (Sigma) and CDR1 protease substrate ([MCA]-Ala-Leu-His-Pro-Glu-Val-Leu-Phe-Val-Leu-Glu-Lys-[DPN][239].

The rate of substrate hydrolysis was monitored for 3 hours by the increase in fluorescence intensity with excitation/emission wavelengths of 328/393 nm for the peptides with MCA/DNP and 335/490nm for those with EDANS/DABCYL, and the relative activity normalized by setting APRc activity as 100%.

SDS-PAGE and Western blotting

SDS-PAGE analysis was performed in a Bio-Rad Mini Protean III electrophoresis apparatus using 4-20% or 12.5% polyacrylamide gels. Samples were treated with loading buffer (0.35 M Tris-HCl, 0.28% SDS buffer pH 6.8, 30% glycerol, 10% SDS, 0.6 M DTT and 0.012% Bromophenol Blue) and boiled for 5 minutes before loading. Gels were stained with Coomassie Brilliant Blue R-250 (Sigma). For Western blot analysis, protein samples were resolved by SDS-PAGE and electrotransferred onto PVDF or nitrocellulose membranes by standard wet (using the buffer 25 mM Tris, 192 mM Glycine and 20% methanol) or semi-dry (in buffer 25 mM Tris, 192 mM Glycine, 20% methanol and 0.025% SDS) transfer apparatus. Membranes were then blocked for one hour in standard TBS containing 1% (v/v) Tween-20 supplemented with 5% (w/v) skim milk or 2% (w/v) BSA and then incubated with the antibody anti-APRc rabbit polyclonal (GenScript, 2 µl/mL). Membranes were washed in TBS containing 0.1% (v/v) Tween-20, incubated with secondary anti-rabbit alkaline phosphatase-conjugated antibodies (GE Healthcare) and revealed using ECF chemiluminescence detection kit (GE Healthcare) in a Molecular Imager FX (Bio-Rad).

Protein quantification

Total protein quantifications were performed either by direct measurement of Abs_{280nm} on a NanoDrop1000 instrument (Thermo Scientific) or using the Pierce BCA Protein Assay Kit or the Bio-Rad Protein Assay Kit (Bradford method), according to the instructions manual. The plates were read in a microplate reader (PowerWave XS Microplate Spectrophotometer, Biotek®).

2.3. Results

Bioinformatics analysis on APRc

In silico analysis of the genome sequence of *R. conorii* str. Malish 7, the etiologic agent of MSF, revealed a gene (*RC1339*) with 696 bp encoding a putative retropepsin-like aspartic protease. This gene is highly conserved among all 55 sequenced *Rickettsia* genomes with deduced amino acid sequences identities ranging from 83.6% (*R. bellii* str. RML369-C APRc homologue) to 100% (e.g., *R. sibirica* 246 APRc homologue) (Table 4). This striking pattern of conservation is illustrated in Figure 9A, which shows the alignment of the deduced amino acid sequence of *R. conorii* RC1339/APRc with eight other homologues from representative species of each rickettsial group (SFG, TG, TRG and AG).

Table 4. Protein sequence identity of each APRc homologue in relation to NP_360976 as well as the taxonomic group of analyzed rickettsial species. Accession numbers are from NCBI database; gene ID from PATRIC database (between parentheses) is also shown from species with the same or non-attributed (NA) accession number. (SFG: spotted fever group; TG: typhus group; TRG: transitional group; AG: ancestral group).

Species	Accession number	Sequence length	Percent Identity	*Rickettsia* Taxonomy
R. conorii **Malish 7**	*NP_360976*	231	-	SFG
R. conorii subsp. indica **ITTR**	*WP_010977893* *(VBIRicCon229600_0066)*	231	100.0	SFG
R. sibirica **246**	*WP_010977893* *(VBIRicSib27963_0845)*	231	100.0	SFG
R. sibirica subsp. mongolitimonae **HA-91**	*WP_010977893* *(VBIRicSib225156_0142)*	231	100.0	SFG
R. sibirica subsp. sibirica BJ-90	*WP_010977893* *(VBIRicSib238733_1447)*	231	100.0	SFG
R. conorii subsp. caspia **A-167**	*WP_016926653*	231	99.6	SFG
R. conorii subsp. israelensis **ISTT CDC1**	*WP_016945366*	231	99.6	SFG
R. parkeri **Portsmouth**	*YP_005393543*	231	99.6	SFG
R. peacockii **Rustic**	*YP_002916969*	231	99.6	SFG
R. rickettsii **Hlp#2**	*WP_01273*	231	99.6	SFG
R. africae **ESF-5**	*YP_002845736*	231	99.1	SFG
R. philipii **364D**	*YP_005301388*	231	99.1	SFG
R. rickettsii **Arizona**	*YP_005289563*	231	99.1	SFG
R. rickettsii **Brazil**	*YP_005294699*	231	99.1	SFG
R. rickettsii **Colombia**	*YP_005288210*	231	99.1	SFG
R. rickettsii **Hauke**	*YP_005293341*	231	99.1	SFG
R. rickettsii **Hino**	*WP_012151442*	231	99.1	SFG
R. slovaca **13-B**	*YP_005066351*	223	99.1	SFG
R. slovaca **D-CWPP**	*WP_014273877*	223	99.1	SFG
R. rickettsii **Sheila Smith**	*YP_001495413*	231	98.8	SFG
R. honei **RB**	*WP_016917263*	231	98.7	SFG

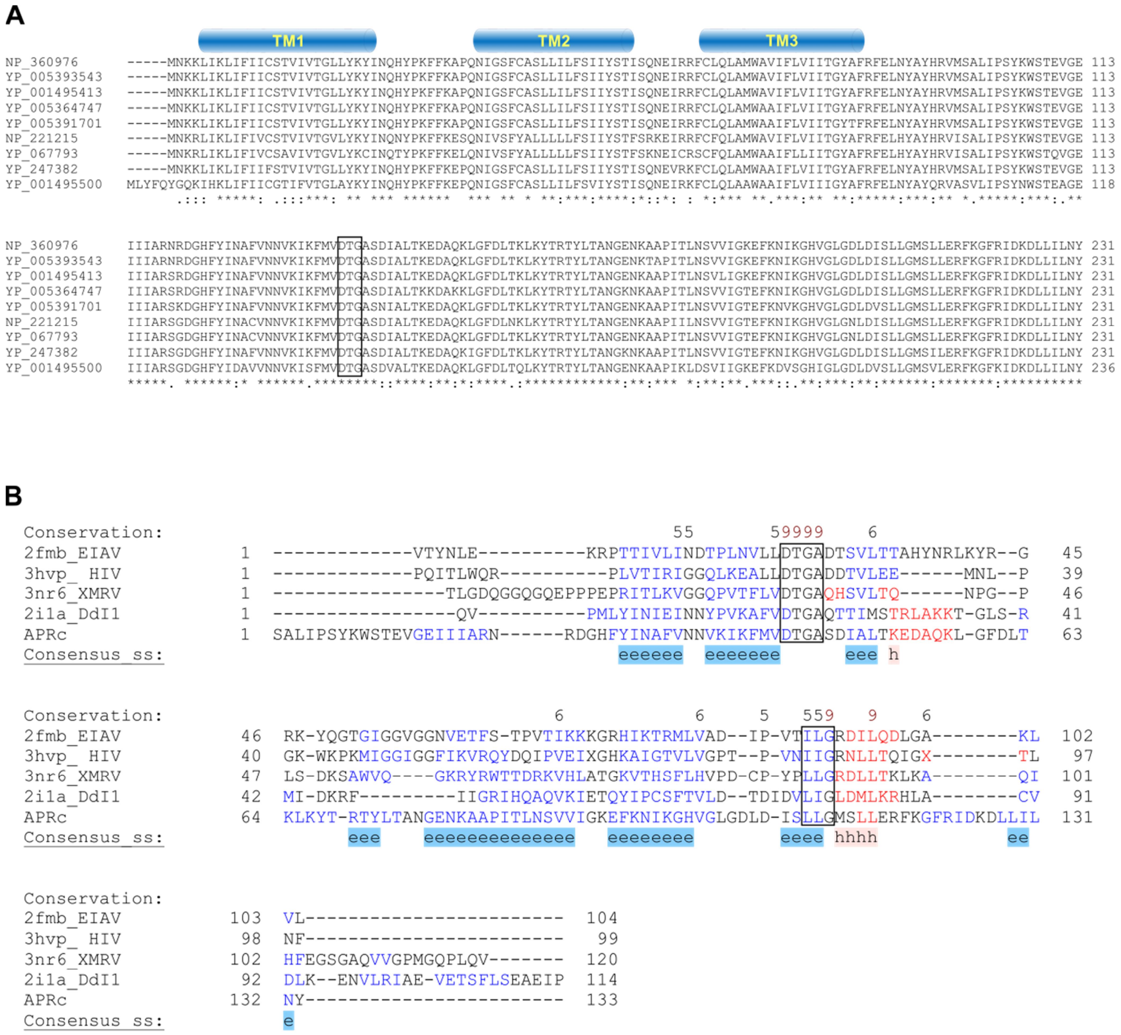
Table 5 (cont.). Protein sequence identity of each APRc homologue in relation to NP_360976 as well as the taxonomic group of analyzed rickettsial species. Accession numbers are from NCBI database; gene ID from PATRIC database (between parentheses) is also shown from species with the same or non-attributed (NA) accession number. (SFG: spotted fever group; TG: typhus group; TRG: transitional group; AG: ancestral group).

Species	Accession number	Sequence length	Percent Identity	*Rickettsia* Taxonomy
***R. rhipicephali* 3-7-female6-CWPP**	*YP_005389841*	231	98.3	SFG
***R. rickettsii* Iowa**	*YP_005286867*	231	98.2	SFG
***R. heilongjiangensis* 054**	*YP_004764976*	231	97.8	SFG
***R. japonica* YH**	*YP_004885318*	231	97.8	SFG
***R. massiliae* AZT80**	*YP_005303600*	231	97.8	SFG
Candidatus *R. amblyommii* GAT-30V	*YP_005364747*	231	97.4	SFG
***R. massiliae* MTU5**	*YP_001499845*	246	97.0	SFG
***R. monacensis* IrR/Munich**	*WP_008580673*	231	97.0	SFG
R.* endosymbiont of *Ixodes scapularis	*WP_008580673*	231	97.0	SFG
***R. montanensis* OSU 85-930**	*YP_005391701*	231	96.5	SFG
***R. australis* Cutlack**	*WP_014412171*	231	95.7	SFG
***R. felis* URRWXCal2**	*YP_247382*	231	95.2	SFG
***R. helvetica* C9P9**	*WP_010420880*	231	94.8	SFG
***R. akari* Hartford**	*WP_012150194*	231	94.4	SFG
***R. prowazekii* Madrid E**	*NP_221215*	231	94.4	TG
***R. prowazekii* Rp22**	*YP_007749515*	231	94.4	TG
***R. prowazekii* Breinl**	*YP_007750927*	231	90.0	TG
***R. prowazekii* BuV67-CWPP**	*YP_005414277*	231	90.0	TG
***R. prowazekii* Cairo 3**	*WP_004596749*	231	90.0	TG
***R. prowazekii* Chernikova**	*YP_005406783*	231	90.0	TG
***R. prowazekii* Dachau**	*YP_005413443*	231	90.0	TG
***R. prowazekii* GvF12**	*WP_004596749*	231	90.0	TG
***R. prowazekii* GvV257**	*YP_005405939*	231	90.0	TG
***R. prowazekii* Katsinyian**	*YP_005407621*	231	90.0	TG
***R. prowazekii* NMRC Madrid E**	*YP_007749515*	231	90.0	TG
***R. prowazekii* RpGvF24**	*YP_005999345*	231	90.0	TG
***R. canadensis* CA410**	*YP_005300059*	231	89.6	TG
***R. canadensis* McKiel**	*YP_001492807*	231	89.2	TG
***R. typhi* B9991CWPP**	*WP_011191285*	231	88.3	TG
***R. typhi* TH1527**	*YP_005424149*	231	88.3	TG
***R. typhi* Wilmington**	*YP_067793*	231	88.3	TG
R. sp. MEAM1 (Bemisia tabaci)	*NA* *(VBIRicSp241202_0036)*	231	85.3	AG
***R. bellii* OSU 85-389**	*YP_001495500*	236	84.0	AG
***R. bellii* RML369-C**	*YP_538487*	236	83.6	AG

A distinguishing feature of rickettsial APs over retroviral-type ones was their predicted membrane-embedded nature, with different algorithms predicting three putative transmembrane α-helix (TMH) segments in the N-terminal domain of APRc (Figure 9A). The presence of Cys residues on these predicted transmembrane regions, which may be linking the three α-helical chains together through interchain disulfide bonds, likely contribute to structural stability. Additionally, an inside orientation for the N terminus and an outside orientation for the C-terminal soluble protease domain of APRc (Arg87-Tyr231) relative to the membrane was predicted by the HMMTOP2[237].

Another striking observation was the apparent absence of sequence homology of this novel type of rickettsial AP when compared with APs from other organisms, except for the presence of the hallmark sequence motifs of family A2 members. Although the overall sequence identity with other retropepsins was found to be lower than 14% (with only 6% for the HIV-1 PR which is considered the archetypal member of this family of APs) it was possible to identify the active site consensus motif Asp-Thr-Gly (contained in the sequence Xaa-Xaa-Asp-Xbb-Gly-Xcc, where a Xaa is hydrophobic, Xbb is Thr or Ser, and Xcc is Ser, Thr or Ala) corresponding to the sequence Met-Val-Asp-Thr-Gly-Ala (amino acids 138-143), followed downstream by a hydrophobic-hydrophobic-Gly sequence (Leu-Leu-Gly, amino acids 208-210). This feature is characteristic of retropepsin-like proteases which are obligate homodimeric enzymes, with each monomer contributing one catalytic triad and one hydrophobic-hydrophobic-Gly motif to form the structural feature known as psi-loop[27,36,41].

An overall retention of structural similarity in proteins highly divergent at the sequence level is often correlated with distant relationships between these proteins. In order to evaluate if this would also apply to APRc, a structure-based alignment of RC1339/APRc soluble catalytic domain with HIV, EIAV and XMRV retropepsins as well as with Ddi1 putative protease domain was performed (Figure 9B). Indeed, this alignment further suggested an overall retention of structural similarity through conservation of the core structural motifs of APs superfamily, despite the high divergence at the sequence level. Importantly, APRc lacks the conserved motif found in most retroviral proteases Gly86-Arg87-Asn/Asp88 (HIV-1 PR numbering), in which the conserved Arg that forms an intra-monomer hydrogen bond with Asp29 in HIV-1 PR is replaced by a Met (Met211).



```
NP_360976       -----MNKKLIKLIFIICSTVIVTGLLYKYINQHYPKFFKAPQNIGSFCASLLILFSIIYSTISQNEIRRFCLQLAMWAVIFLVIITGYAFRFELNYAYHRVMSALIPSYKWSTEVGE 113
YP_005393543    -----MNKKLIKLIFIICSTVIVTGLLYKYINQHYPKFFKAPQNIGSFCASLLILFSIIYSTISQNEIRRFCLQLAMWAVIFLVIITGYAFRFELNYAYHRVMSALIPSYKWSTEVGE 113
YP_001495413    -----MNKKLIKLIFIICSTVIVTGLLYKYINQHYPKFFKAPQNIGSFCASLLILFSIIYSTISQNEIRRFCLQLAMWAVIFLVIITGYAFRFELNYAYHRVMSALIPSYKWSTEVGE 113
YP_005364747    -----MNKKLIKLIFIICSTVIVTGLLYKYINQHYPKFFKAPQNIGSFCASLLILFSIIYSTISQNEIRRFCLQLAMWAAIFLVIITGYAFRFELNYAYHRVMSALIPSYKWSTEVGE 113
YP_005391701    -----MNKKLIKLIFIICSTVIVTGLLYKYINQHYPKFFKAPQNIGSFCASLLILFSIIYSTISQNEIRRFCLQLAMWAAIFLVIITGYTFRFELNYAYHRVMSALIPSYKWSTEVGE 113
NP_221215       -----MNKRLIKLIFIVCSTVIVTGVLYKYINQNYPKFFKESQNIVSFYALLLLLFSIIYSTFSRKEIRRFCFQLAMWAVIFLVIITGYAFRFELHYAYHRVISALIPSYKWSTEVGE 113
YP_067793       -----MNKRLIKLIFIVCSAVIVTGVLYKCINQTYPKFFKELQNIVSFYALLLLLFSIIYSTFSKNEICRSCFQLAMWAAIFLLIITGYAFRFELHYAYHRVISALIPSYKWSTQVGE 113
YP_247382       -----MNKKLIKLIFIIFSTVIVTGLLYKYINQHYPKFFKEPQNIGSFCASLLILFSIIYSTISQNEVRKFCLQLAMWAAIFLVIITGYAFRFELNYAYHRVMSALIPSYKWSTEVGE 113
YP_001495500    MLYFQYGQKIHKLIFIICGTIFVTGLAYKYINQHYPKFFKEPQNIGSFCASLLILFSVIYSTISQNEIRKFCLQLAAWAAIFLVIIIGYAFRFELNYAYQRVASVLIPSYNWSTEAGE 118
                      .::: *****: .:::***: ** *** ******  *** ** * **:***:****:*::*: : *:*** **.***:** **:*****:***:** *.*****:***:.**

NP_360976       IIIARNRDGHFYINAFVNNVKIKFMVDTGASDIALTKEDAQKLGFDLTKLKYTRTYLTANGENKAAPITLNSVVIGKEFKNIKGHVGLGDLDISLLGMSLLERFKGFRIDKDLLILNY 231
YP_005393543    IIIARNRDGHFYINAFVNNVKIKFMVDTGASDIALTKEDAQKLGFDLTKLKYTRTYLTANGENKTAPITLNSVVIGKEFKNIKGHVGLGDLDISLLGMSLLERFKGFRIDKDLLILNY 231
YP_001495413    IIIARSRDGHFYINAFVNNVKIKFMVDTGASDIALTKEDAQKLGFDLTKLKYTRTYLTANGENKAAPITLNSVLIGKEFKNIKGHVGLGDLDISLLGMSLLERFKGFRIDKDLLILNY 231
YP_005364747    IIIARSRDGHFYINAFVNNVKIKFMVDTGASDIALTKKDAKKLGFDLTKLKYTRTYLTANGKNKAAPITLNSVVIGTEFKNIKGHVGLGDLDISLLGMSLLERFKGFRIDKDLLILNY 231
YP_005391701    IIIARSKDGHFYINAFVNNVKIKFMVDTGASNIALTKEDAQKLGFDLTKLKYTRTYLTANGENKAAPITLNSVVIGTEFKNVKGHVGLGDLDVSLLGMSLLERFKGFRIDKDLLILNY 231
NP_221215       IIIARSGDGHFYINACVNNVKIKFMVDTGASDIALTKEDAQKLGFDLNKLKYTRTYLTANGENKAAPIILNSVVIGTEFKNIKGHVGLGNLDISLLGMSLLERFKGFRIDKDLLILNY 231
YP_067793       IIIARSGDGHFYINACVNNVKIKFMVDTGASDIALTKEDAQKLGFDLTKLKYTRTYLTANGENKAAPITLNSVVIGTEFKNIKGHVGLGNLDISLLGMSLLERFKGFRIDKDLLILNY 231
YP_247382       IIIARSGDGHFYINAFVNNVKIKFMVDTGASDIALTKEDAQKIGFDLTKLKYTRTYLTANGKNKAAPITLNSVVIGTEFKNIKGHVGLGDLDISLLGMSILERFKGFRIDKDLLILNY 231
YP_001495500    IIIARSGDGHFYIDAVVNNVKISFMVDTGASDVALTKEDAQKLGFDLTQLKYTRTYLTANGENKAAPIKLDSVIIGKEFKDVSGHIGLGDLDVSLLGMSVLERFKGFKIDKDLLILNY 236
                *****. ******:* ******.********::****:**:*:****.:**********:**:*** *:**:**.***::.**:***:**:******:*******:**********
```

```
Conservation:                                              55       59999    6
2fmb_EIAV          1  -------------VTYNLE----------KRPTTIVLINDTPLNVLLDTGADTSVLTTAHYNRLKYR--G   45
3hvp_ HIV          1  -------------PQITLWQR----------PLVTIRIGGQLKEALLDTGADDTVLEE------MNL--P   39
3nr6_XMRV          1  ----------------TLGDQGGQGQEPPPEPRITLKVGGQPVTFLVDTGAQHSVLTQ------NPG--P   46
2i1a_DdI1          1  -----------------QV----------PMLYINIEINNYPVKAFVDTGAQTTIMSTRLAKKT-GLS-R   41
APRc               1  SALIPSYKWSTEVGEIIIARN------RDGHFYINAFVNNVKIKFMVDTGASDIALTKEDAQKL-GFDLT   63
Consensus_ss:                                         eeeeee  eeeeeee      eee h

Conservation:                                     6            6     5   559   9    6
2fmb_EIAV         46  RK-YQGTGIGGVGGNVETFS-TPVTIKKKGRHIKTRMLVAD--IP-VTILGRDILQDLGA--------KL  102
3hvp_ HIV         40  GK-WKPKMIGGIGGFIKVRQYDQIPVEIXGHKAIGTVLVGPT--P-VNIIGRNLLTQIGX--------TL   97
3nr6_XMRV         47  LS-DKSAWVQ----GKRYRWTTDRKVHLATGKVTHSFLHVPD-CP-YPLLGRDLLTKLKA--------QI  101
2i1a_DdI1         42  MI-DKRF---------IIGRIHQAQVKIETQYIPCSFTVLD--TDIDVLIGLDMLKRHLA--------CV   91
APRc              64  KLKYT-RTYLTANGENKAAPITLNSVVIGKEFKNIKGHVGLGDLD-ISLLGMSLLERFKGFRIDKDLLIL  131
Consensus_ss:                eee    eeeeeeeeeeeeee   eeeeeeee        eeee hhhh            ee

Conservation:
2fmb_EIAV        103  VL------------------------  104
3hvp_ HIV         98  NF------------------------   99
3nr6_XMRV        102  HFEGSGAQVVGPMGQPLQV-------  120
2i1a_DdI1         92  DLK--ENVLRIAE-VETSFLSEAEIP  114
APRc             132  NY------------------------  133
Consensus_ss:         e
```

Figure 9. *Pattern of sequence conservation among RC1339/APRc gene homologues and structural similarity of APRc with other retropepsin enzymes.* (**A**) Multi-alignment of deduced amino acid sequences of the putative retropepsin-like protease from representative species from all rickettsial taxonomic groups (spotted fever group, typhus group, transitional group and ancestral group). Sequences were aligned against RC1339/APRc sequence from R. conorii (NP_360976) using the ClustalW software[235]. Accession numbers and corresponding species are described in Table 4. The predicted α-helical transmembrane domains are represented by cylinders and the box indicates the active site motif (DTG). (**B**) Structure-based alignment of the soluble catalytic domain of RC1339/APRc with HIV-1 (PDB 3hvp), EIAV (PDB 2fmb) and XMRV (PDB 3nr6) retropepsins and with DdI1 putative protease domain (PDB 2i1a), performed with PROMALS3D[238]. The first line shows conservation indices for positions with a conservation index above 4. Consensus_ss represent consensus predicted secondary structures (alpha-helix: h; beta-strand: e). Sequences are colored according to predicted secondary structures (red: alpha-helix, blue: beta-strand). Red nines highlight the most conserved positions. Active site consensus motif Asp-Thr-Gly and hydrophobic-hydrophobic-Gly sequence are boxed.

APRc autoprocessing activity and dependence on the catalytic aspartate residue

The lack of effective and easy tools for site-specific gene inactivation and complementation in *Rickettsia* has led to the use of heterologous expression in surrogate hosts to explore different aspects of gene activity, function or regulation[193,240,241]. In this work, we followed a similar strategy to conduct comprehensive biochemical and enzymatic characterization studies on APRc. Using *R. conorii* RC1339 as our working model, we first sought to investigate whether the putative gene would in fact encode a functional and active AP by producing its soluble catalytic domain fused to GST (rGST-APRc$_{87-231}$) in *E. coli*. Assuming the predicted boundary between the transmembrane and soluble catalytic domains at Phe86-Arg87, the synthetic codon optimized sequence coding for the whole soluble domain was cloned into pGEX-4T2 (pGST-APRc$_{87-231}$) and the fusion construct expressed in *E. coli* (BL21 Star (DE3) strain).

To purify recombinant rGST-APRc$_{87-231}$, the soluble fraction of the cell lysates was applied to a GSTrap HP affinity chromatography (Figure 10A) and the eluted fractions were pooled and further purified by size-exclusion chromatography on a Superdex 200 HiLoad 26/60 (Figure 10B). As shown in Figure 10C, purified fractions analyzed by SDS-PAGE confirmed the presence of the fusion protein with approximately 42 KDa as well as free GST (25 kDa). In parallel, the same purification protocol was used to purify the active site mutant rGST-APRc(D140A)$_{87-231}$, where the putative catalytic aspartate residue was mutated to an alanine, and the same pattern of purification was observed (Figure 10D). These results suggest that the considerable amounts of free fusion tag likely result from proteolytic degradation by the host and not from autolytic activity of APRc. Since GST forms dimers under normal purification conditions[242], free GST may form stable dimers with rGST-APRc$_{87-231}$ fusion protein which can explain the unsuccessful attempts to improve purity of these samples.

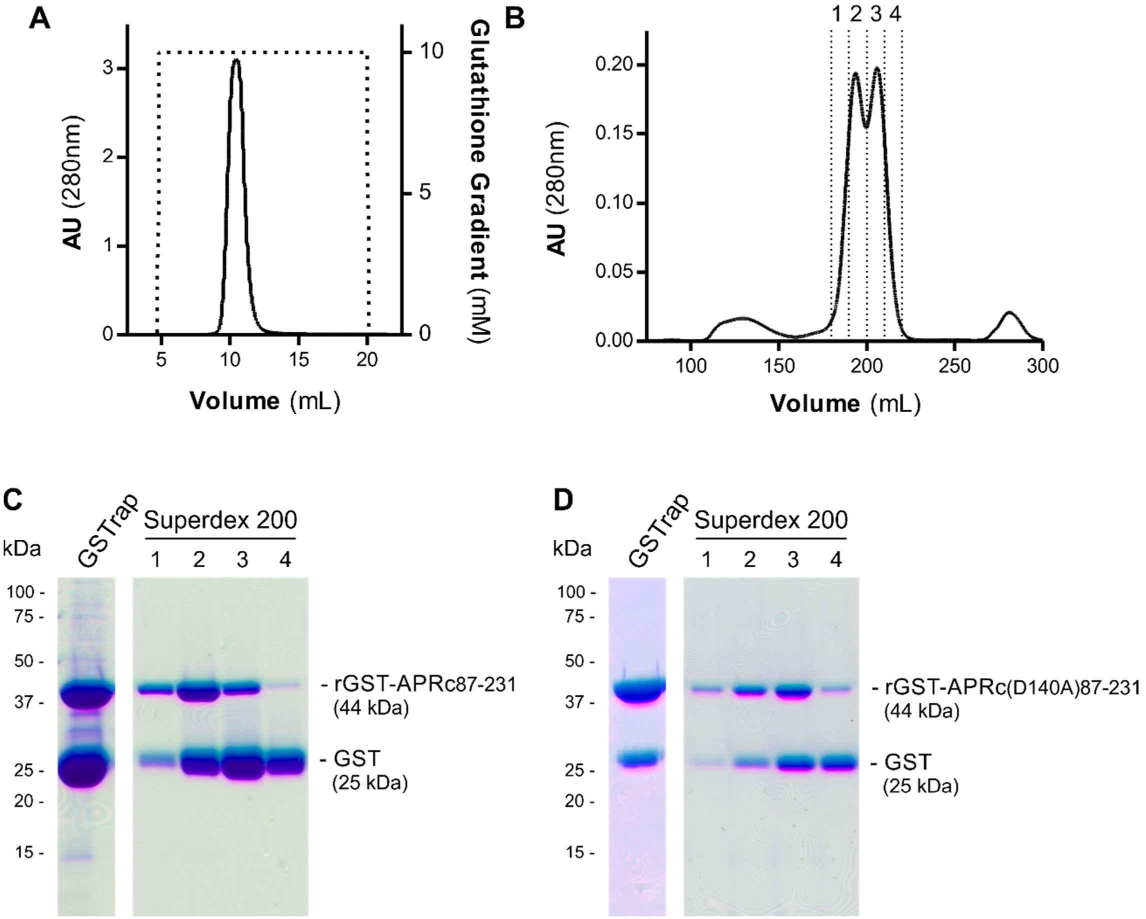


Figure 10. *Purification of rGST-APRc$_{87-231}$ and its active site mutant by affinity and size-exclusion chromatographies.* (**A**) Total soluble extracts of *E. coli* overexpressing rGST-APRc$_{87-231}$ were loaded on a GSTrap 5 mL column previously equilibrated with 50 mM Tris-HCl pH 7.5 at a flow rate of 2 mL/min and elution carried out with PBS, 10 mM Glutathione. Dotted lines in the chromatogram represent the glutathione gradient. (**B**) The eluted protein was then applied to a Superdex 200 HiLoad 26/60 column equilibrated with PBS buffer at a flow rate of 2 mL/min. Protein elution was monitored by measuring the A_{280nm}. (**C**) Fractions with eluted protein outlined by dotted lines in the chromatogram were analyzed in a SDS-PAGE gel stained with Coomassie blue. In addition to rGST-APRc$_{87-231}$ precursor form, fractions 1-4 from the size exclusion chromatography (SEC) exhibited a high degree of contamination with free GST. (**D**) When the same strategy of purification was applied to the active site mutant rGST-APRc(D140A)$_{87-231}$, an identical pattern of purification was observed. Description of the recombinant proteins are indicated on the right side of the gel and molecular weight markers in kilodaltons (kDa) are shown on the left.

One shared feature of proteases of the retropepsin family is their autoprocessing activity which promotes their own release from a larger polyprotein precursor[27]. In order to start assessing whether APRc also displays this activity, the recombinant fusion protein was incubated under different pH values (3.0 - 7.0) for about 24 h. As shown in Figure 11A, our results demonstrate that recombinant rGST-APRc$_{87-231}$ also undergone autoprocessing *in vitro*, with optimal activity at pH 6.0, as indicated by the higher accumulation of a cleavage product with the lowest molecular weight (denoted by * in Figure 11A). Interestingly, this autolytic activity was shown to be a multi-step process given by the sequential generation of three cleavage products over a 48 h time course (Figure 11B, left panel). Edman sequencing of these APRc fragments allowed the identification of the three autolytic cleavage sites: Tyr92-Ala93, Met98-Ser99 and Ser104-Tyr105, as depicted Figure 11C.

In order to evaluate the role of the putative catalytic aspartate for this autoprocessing activity, the produced active site mutant of rGST-APRc$_{87-231}$, was activated under the same conditions as those used for the wild-type fusion protein. As expected, the mutation significantly affected the activation process (Figure 11B, right panel), suggesting that APRc is dependent on the conserved catalytic aspartate residue for triggering autolytic activity.

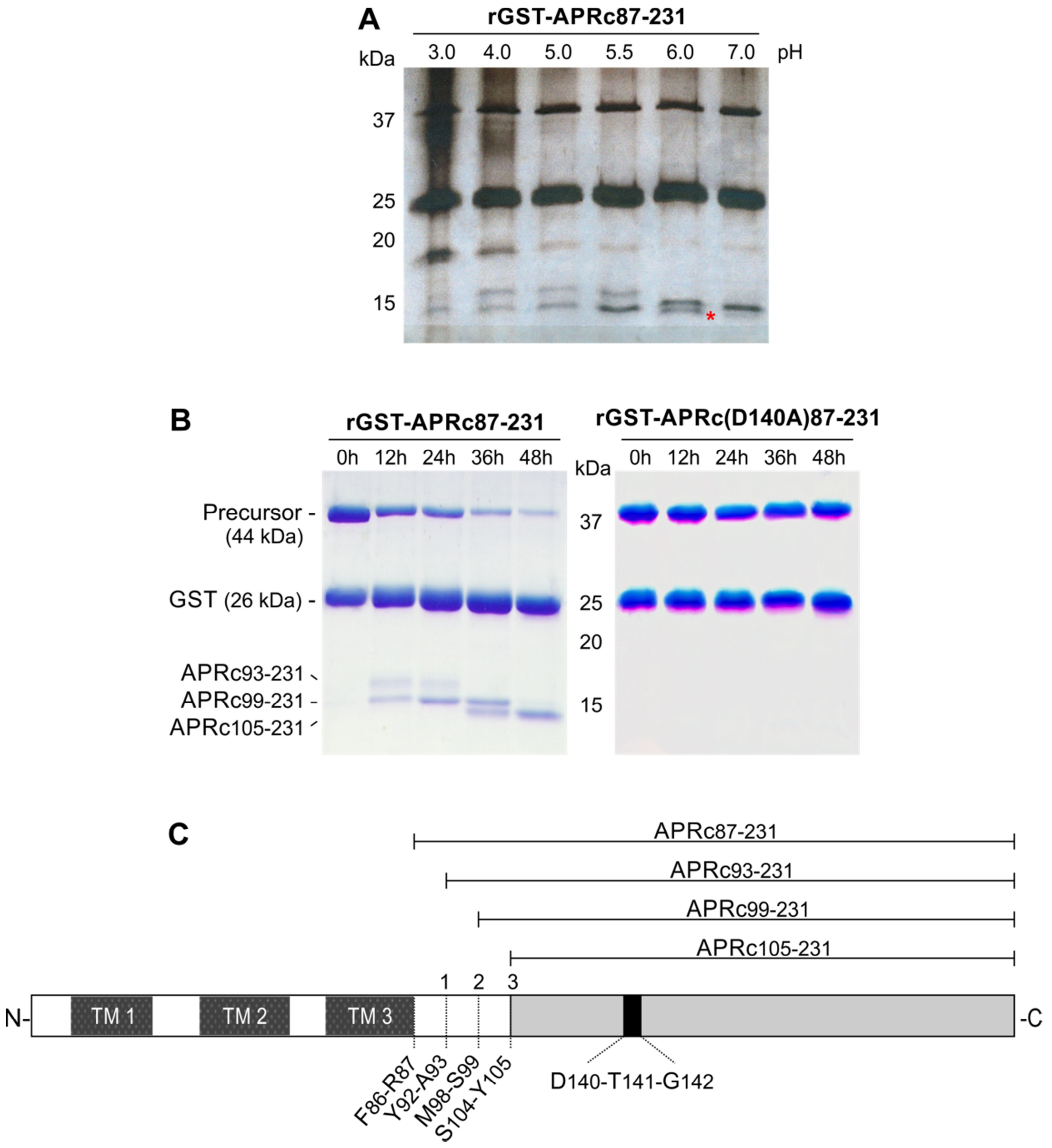


Figure 11. *Autoprocessing activity of recombinant APRc soluble catalytic domain and dependence on the catalytic aspartate residue.* (**A**) The soluble catalytic domain (amino acids 87-231) was fused to GST and produced in *E. coli*. Upon purification, the auto-activation of rGST-APRc$_{87-231}$ was first evaluated *in vitro* at different pH values (3-7). APRc samples were diluted 1:1 with 0.1 M sodium citrate buffer pH 3, 0.1 M sodium acetate buffer pH 4, pH 5, pH 5.5, pH 6 and 0.1 M Tris-HCl pH 7, and incubated at 37 ˚C for approximately 24 h. The analysis by SDS-PAGE (stained with silver nitrate) revealed an optimal auto-processing activity at pH 6, as indicated by faster generation of the cleavage product with the lowest molecular weight (denoted by *). (**B**) Subsequent auto-activation studies of rGST-APRc$_{87-231}$ were performed in 0.1 M sodium acetate buffer pH 6 at 37 ˚C for 48 h and monitored by SDS-PAGE stained with Coomassie blue. rGST-APRc$_{87-231}$ undergoes multi-step auto-activation processing, resulting in the formation of the activated form APRc$_{105-231}$-His with ~14.2 kDa (left panel). Mutation of the active site aspartic acid by alanine in this fusion construct [rGST-APRc(D140A) $_{87-231}$] clearly impaired the auto-catalytic activity of the protease (right panel). (**C**) Schematic representation of full-length APRc domain organization. APRc is predicted to comprise three transmembrane domains (TM 1-3) at the N terminus and the soluble catalytic domain at the C terminus. The three auto-cleavage sites (shown in B) identified by Edman degradation are highlighted by order of cleavage (1-3). Incubation time course in hours (h) are indicated above gels and the molecular weight markers in kilodaltons (kDa) are shown on the left.

As a first approach to assess enzyme activity, we used oxidized insulin β chain as a substrate as this polypeptide is usually cleaved by APs, and tested its cleavage over activation time for purified rGST-APRc$_{87\text{-}231}$. As illustrated in Figure 12A, samples from each time point (0, 12, 24, 36 and 48 h) were tested and the reaction products separated by RP-HPLC. Interestingly, the presence of several insulin cleavage products was concomitant with the appearance of the activation product APRc$_{105\text{-}231}$-His, suggesting that autoprocessing may be an essential step for the activation of recombinant APRc. This was also confirmed by the lack of activity of the active site mutant rGST-APRc(D140A)$_{87\text{-}231}$ towards insulin β chain, after incubation for 48 h at pH 6.0 (comparison in Figure 12B).

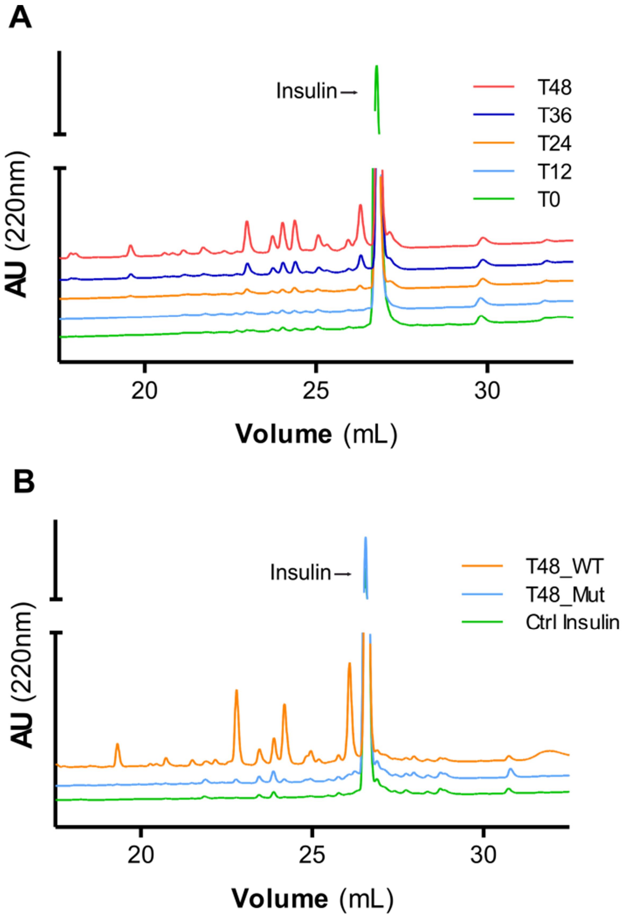


Figure 12. *Oxidized insulin β chain degradation by APRc.* (**A**) Activity of rGST-APRc$_{87\text{-}231}$ towards oxidized insulin β chain was tested over activation time. Samples corresponding to the different time points of activation, T0, T12, T24, T36 and T48 h, were incubated with the substrate at pH 6.0 for 16 h. Reaction products were then evaluated by RP-HPLC showing that substrate cleavage (appearance of four major peaks) was concomitant with appearance of the final activation product. (**B**) The activity of wild-type rGST-APRc$_{87\text{-}231}$ towards oxidized insulin β chain upon activation assays *in vitro* for 48 h was compared to that of rGST-APRc(D140A)$_{87\text{-}231}$. T48_WT and T48_Mut correspond to the analysis of reaction products by RP-HPLC for the wild-type and active site mutant, respectively. Ctrl Insulin corresponds to the RP-HPLC profile of oxidized insulin β chain in the absence of protease. The observation of several peaks corresponding to insulin cleavage products only upon incubation with wild-type protease, confirmed that these peaks resulted from APRc activity.

The observed autoprocessing ability of APRc and the importance of the catalytic aspartate were further confirmed by expressing the constructs harboring the soluble domain (rGST-APRc$_{87\text{-}231}$) and its active site mutant (rGST-APRc(D140A)$_{87\text{-}231}$) in *E. coli* and by analyzing total soluble fractions for the presence of APRc activated forms with a specific APRc polyclonal antibody (raised towards amino acids 165-178). As shown in Figure 13, and consistent with the results obtained in our *in vitro* assays, the activation products were only detected when the wild-type sequence was expressed, further corroborating the role of the catalytic aspartate in this autoprocessing activity. It is also noteworthy that this activation occurred at a much higher rate than what was observed *in vitro*, indicating that the *E. coli* cytoplasm probably offers more suitable conditions for APRc processing. This intrinsic autoprocessing observed during expression in *E. coli* is in line with what has been documented for other retropepsin-type APs (e.g., HIV-1 and XMRV PRs)[65,106].

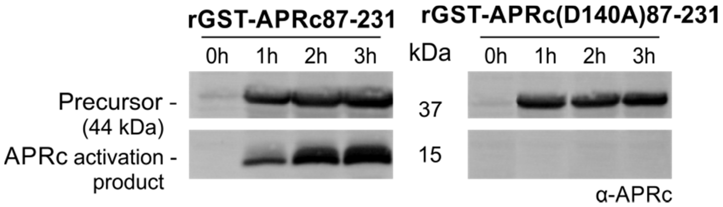


Figure 13. *Immunoblot detection of rGST-APRc$_{87\text{-}231}$ /rGST-APRc(D140A)$_{87\text{-}231}$ over expression time.* APRc auto-processing ability was evaluated in total lysates of *E. coli* cells overexpressing wild-type rGST-APRc$_{87\text{-}231}$ or the correspondent active site mutant rGST-APRc(D140A)$_{87\text{-}231}$ over a time-course of 3 h and subsequently subjected to Western blot analysis with anti-APRc antibody. A band with approximately 15 kDa was only detected for the wild-type construct. Expression time course in hours (h) is indicated above gels and the molecular weight markers in kilodaltons (kDa) are shown on the left.

Auto-processing studies on the last intermediate of APRc activation

Since the expression of APRc soluble domain fused to GST resulted in a high degree of contamination with free GST, an alternative strategy was undertaken to streamline the production of APRc activation product with higher yield and purity. For this we designed a new construct where the sequence encoding the intermediate of activation APRc$_{99-231}$ was cloned into pET23a expression vector (Invitrogen) in frame with a C-terminal 6xHis-tag (rAPRc$_{99-231}$-His). This construct was readily expressed in the soluble form in *E. coli* and a purification protocol was optimized consisting of a Ni-IMAC step, followed by dialysis of APRc-enriched polled fractions, and further purification through a cation exchange chromatography with a Mono S GL column (Figure 14). From the SDS-PAGE analysis (Figure 14C) it is clear that a highly expressed protein with an apparent MW of 16 kDa is purified on the 100 mM Imidazole elution step of Ni-IMAC chromatography (Figure 14A). In addition, one can also observe the presence of a lower molecular weight band with approximately 15 kDa, likely due to proteolytic processing occurring during the heterologous expression in *E. coli* cells. When this fraction is subjected to cation exchange purification, three major peaks are obtained (Figure 14B) corresponding to different proportions of these two forms, as it is clearly observed on SDS-PAGE analysis (Figure 14C). Given the consistent presence of both products throughout different purification batches (with slight variations on the amount of the minor processing product), eluted fractions from the Mono S column were always combined in a single pool after purification for subsequent assays.

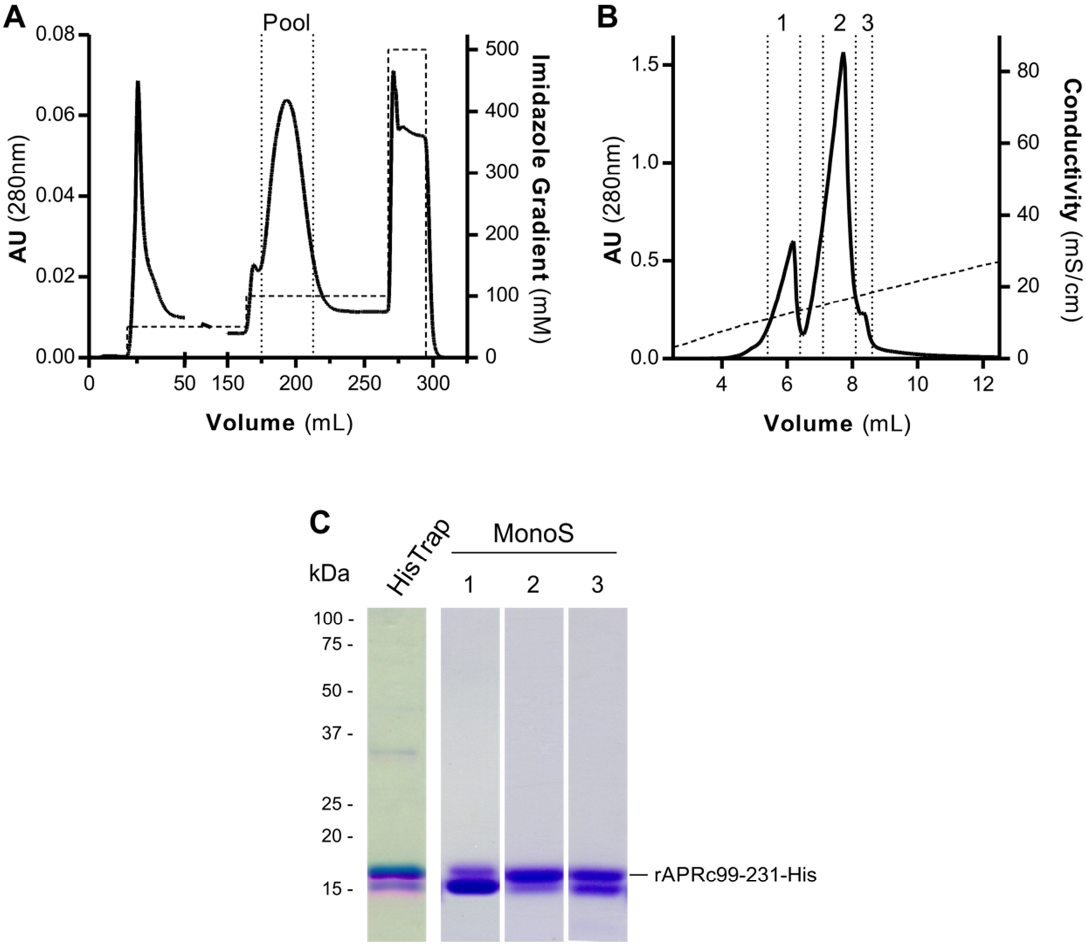


Figure 14. *Purification of rAPRc$_{99\text{-}231}$-His by affinity and cation exchange chromatographies.* (**A**) Recombinant rAPRc$_{99\text{-}231}$-His was first purified on a HisTrap 5 mL column pre-charged with Ni^{2+} ions and equilibrated in 20 mM phosphate buffer pH 7.5, 500 mM NaCl and 10 mM Imidazole. The protein was eluted with a three-step gradient of 50 mM, 100 mM and 500 mM Imidazole at a flow rate of 2 mL/min and monitored by measuring the A$_{280nm}$. The eluted protein from the 100 mM Imidazole gradient step outlined by dotted lines was pooled, dialyzed for 16 h towards 20 mM Phosphate buffer pH 7.5 and loaded on a (**B**) MonoS 5/50 GL column equilibrated in the same buffer. Protein elution was carried out by a continuous salt gradient (0-1 M NaCl), at a flow rate of 0.75 mL/min and monitored by measuring the A$_{280nm}$. (**C**) Approximately 4 μg of protein from the HisTrap column and from fractions 1, 2 and 3 outlined by dotted lines in the chromatograms were analyzed in a SDS-PAGE gel stained with Coomassie blue. The two purified forms of rAPRc$_{99\text{-}231}$-His corresponding to the precursor and processed form are indicated in the gel. The molecular weight markers in kilodaltons (kDa) are shown on the left.

To further substantiate that the observed lower molecular weight product was a result of APRc autoprocessing activity (as previously observed for the GST-fusion protein), three additional mutated forms of the intermediate of activation rAPRc$_{99\text{-}231}$-His were generated. Two of the mutants comprise the substitution of either P1 or P1' residues from the last cleavage site Ser104*Tyr105 (* denotes cleavage site 3, Figure 11C) by a proline: rAPRc(S104P)$_{99\text{-}231}$-His and rAPRc(Y105P)$_{99\text{-}231}$-His), whereas the third construct corresponds to the active site mutant (rAPRc(D140A)$_{99\text{-}231}$-His). By changing the cleavage site or abolishing enzyme activity, a significant impact on autoprocessing ability was expected. These proteins were expressed and purified under the same experimental conditions used for rAPRc$_{99\text{-}231}$-His. Although the same protein level and purity were obtained as with rAPRc$_{99\text{-}231}$-His, major differences on the content of purified samples were identified in what concerns the presence of processed forms (immediately visible also by the absence of the first peak in the MonoS chromatogram). The analysis of purified APRc mutants for the last cleavage site rAPRc(S104P)$_{99\text{-}231}$-His (Figure 15A) and rAPRc(Y105P)$_{99\text{-}231}$-His (Figure 15B), revealed that the processing of APRc into the final form appears to be tightly regulated in a sequence-specific manner, since only a negligible amount of the low molecular weight band was observed for rAPRc(S104P)$_{99\text{-}231}$-His (Figure 15A, fraction 1), with no visible product in the Y105P construct . Finally, the absence of processed forms (Figure 15C) in the active site mutant rAPRc(D140A)$_{99\text{-}231}$-His confirmed again the critical role of the catalytic aspartate on the maturation process of this AP.

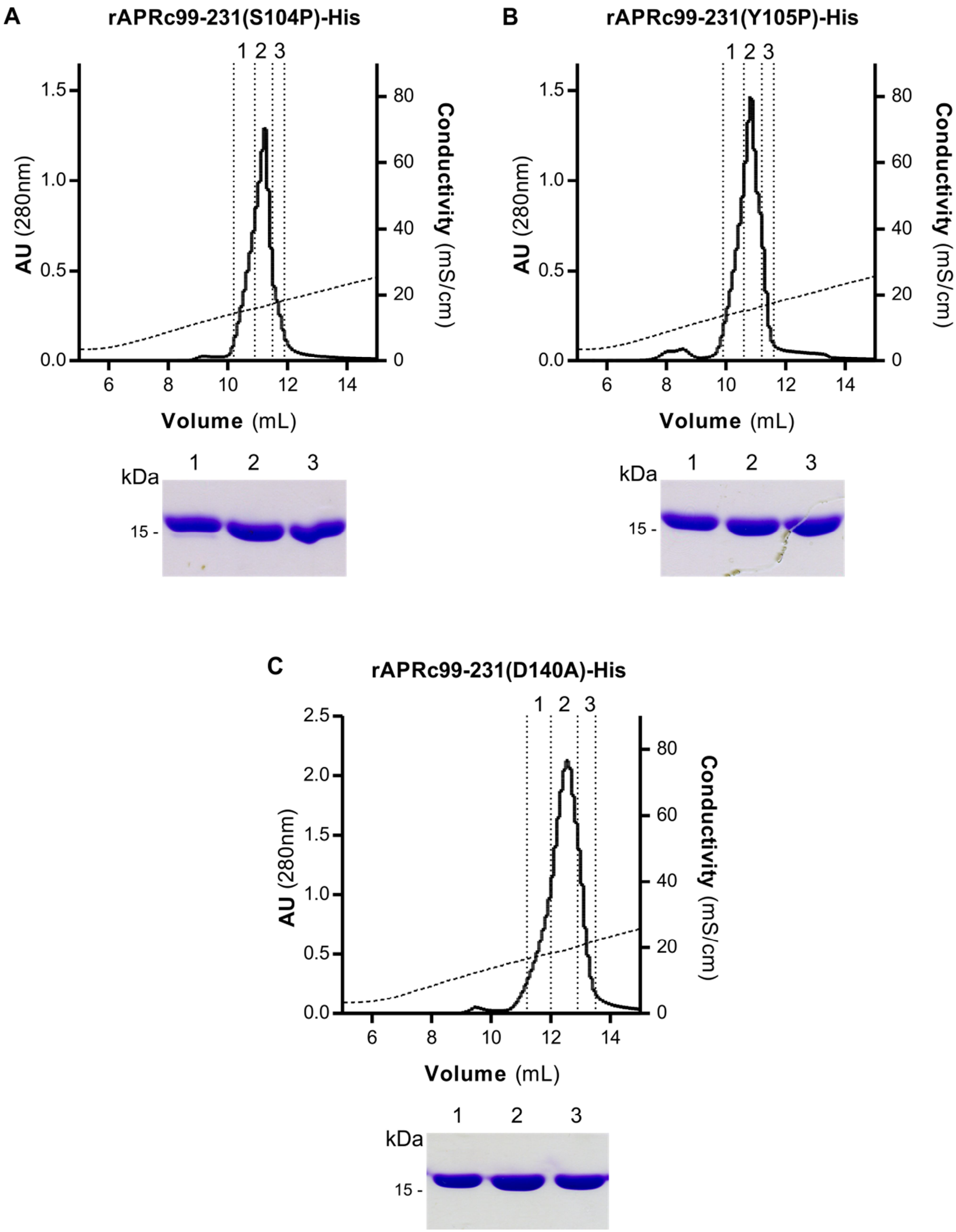


Figure 15. *Purification of rAPRc$_{99\text{-}231}$-His mutants by cation exchange chromatography.* Recombinant rAPRc$_{99\text{-}231}$-His mutants were first purified on a HisTrap 5 mL column as described for wild-type protease in the legend of Figure 14. (**A**) rAPRc$_{99\text{-}231}$(S104P)-His, (**B**) rAPRc$_{99\text{-}231}$(Y105P)-His and (**C**) rAPRc$_{99\text{-}231}$(D140A)-His elution on MonoS 5/50 GL column was carried out by a continuous salt gradient (0-1 M NaCl), at a flow rate of 0.75 mL/min and monitored by measuring the A$_{280nm}$. Approximately 4 μg of protein from fractions 1, 2 and 3 outlined by dotted lines in each chromatogram were analyzed by SDS-PAGE stained with Coomassie blue. The molecular weight markers in kilodaltons (kDa) are shown on the left.

To determine whether these differences were the result of an effect of mutations on protein structural stability, all four constructs were analyzed by circular dichroism (CD) at

Applied Photophysics. The analyses of far-UV CD spectra obtained for each construct have shown that they share the same core secondary structures, thereby reflecting an identical fold pattern (Table 6).

Table 6. *Circular dichroism analyses of rAPRc$_{99-231}$-His and corresponding mutants*. The percentages of α-helix, β-strands and coils were obtained by deconvolution of the experimental far-UV spectra (195-260 nm) using the Net33 model (CDNN CD Spectra Deconvolution Software[243]).

	Alpha-Helix	Beta-Antiparallel	Beta-Parallel	Beta-Turn	Random Coil
rAPRc$_{99-231}$-His	4.4 – 9.4 %	39.3 – 42.5 %	5.3 – 6.0 %	17.4 – 19.4 %	31.0 – 35.3 %
rAPRc(D140A)$_{99-231}$-His	4.4 – 9.3 %	39.0 – 42.2 %	5.3 – 5.9 %	17.6 – 19.5 %	31.2 – 35.4 %
rAPRc(S104P)$_{99-231}$-His	4.4 – 9.3 %	39.3 – 42.6 %	5.3 – 6.0 %	17.4 – 19.5 %	31.0 – 35.3 %
rAPRc(Y105P)$_{99-231}$-His	4.3 – 8.9 %	39.6 – 42.8 %	5.3 – 5.8 %	17.7 – 19.4 %	31.5 – 35.5 %

In contrast, when the CD spectrum for each protein was analyzed as a function of temperature to determine the midpoint of the unfolding transition (T_M), a major difference in protein stability was observed for the construct rAPRc(D140A)$_{99-231}$-His. Accordingly, while the wild-type construct rAPRc$_{99-231}$-His and the mutated constructs rAPRc(S104P)$_{99-231}$-His and rAPRc(Y105P)$_{99-231}$-His exhibited identical T_M of 49.2±1.2, 50.0±0.2 and 51.8±0.3 ˚C, respectively, a significant decrease of the T_M (43.0±0.5 ˚C) was observed for the rAPRc(D140A)$_{99-231}$-His construct. Taken together, these results show that only the mutation of the catalytic aspartate (D140A) has an impact on the stability of APRc, apparently without affecting the overall protein folding. This result is not totally unexpected, as the catalytic aspartate has been shown to play a significant role not only for catalytic activity but also in stabilizing the monomer and dimer folds of retropepsins without significantly changing the protein structure[49,244,245].

Despite the presence of some processed product upon purification of wild-type rAPRc$_{99-231}$-His, we wanted to evaluate if, as shown for rGST-APRc$_{87-231}$, this intermediate of activation would also be able to undergo auto-activation *in vitro* into the mature form at pH 6.0. The results in Figure 16A confirm this by showing protein conversion over activation time. To further characterize APRc enzymatic activity we designed a specific fluorogenic substrate which mimics the identified auto-cleavage site between Ser104-Tyr105 residues (PepRick14 peptide: MCA-Lys-Ala-Leu-Ile-Pro-Ser-Tyr-Lys-Trp-Ser-Lys-DNP) and tested this substrate during rAPRc$_{99-231}$-His activation. As previously observed with the GST-fusion precursor, activity towards this substrate was shown to be also dependent on the conversion step and the highest activity observed upon accumulation of the conversion product (Figure 16A-B), further strengthening the importance of enzyme activation.

Given the observed impact of mutating the catalytic Asp residue for APRc autoprocessing ability we decided to evaluate the effect of pepstatin (the classical inhibitor of aspartic proteases) and indinavir (an HIV-1 PR inhibitor) in rAPRc$_{99-231}$-His autoprocessing. Our results (Figure 16C) show that in the presence of pepstatin the auto-activation step was slightly slowed whereas indinavir had no apparent inhibitory effect on this autoprocessing activity. These results are not completely unexpected, as studies on HIV-1 PR precursor maturation have also shown different susceptibilities of autoprocessing to pepstatin and indinavir[65,246]. Surprisingly, EDTA inhibited rAPRc$_{99-231}$-His auto-activation suggesting that a metal ion may be involved in proper folding and/or enzyme activity.

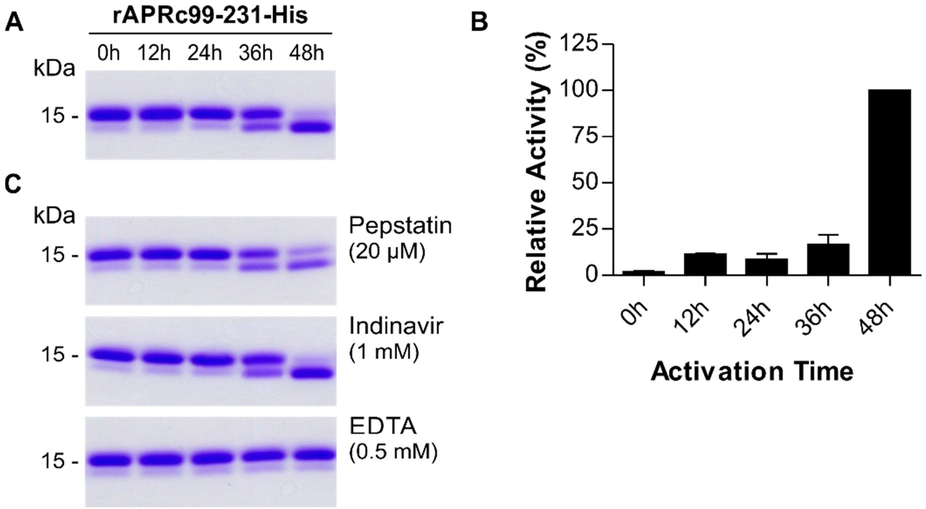


Figure 16. *Auto-processing activity of the last intermediate of activation rAPRc$_{99-231}$-His***.** (**A**) The intermediate of activation rAPRc$_{99-231}$-His was fused to C-terminal His-tag and produced in *E. coli*. Upon purification, the auto-activation assays were performed *in vitro* in 0.1 M sodium acetate buffer pH 6 at 37 °C for 48h and monitored by SDS-PAGE stained with Coomassie blue. rAPRc$_{99-231}$-His undergone auto-processing, resulting in the formation of the activated form. (**B**) Activity of rAPRc$_{99-231}$-His towards the fluorogenic substrate MCA-Lys-Ala-Leu-Ile-Pro-Ser-Tyr-Lys-Trp-Ser-Lys-DNP was tested over activation time. Substrate cleavage increased with accumulation of the final activation product. The error bars represent standard deviation of the mean. (**C**) When the auto-activation studies were carried out in the presence of pepstatin, this conversion was slower and no significant effect was detected under the presence of indinavir. The presence of EDTA completely inhibited protease conversion. The molecular weight markers in kilodaltons (kDa) are shown on the left.

Interestingly, when the final product APRc$_{105-231}$ was directly produced in *E. coli* with a C-terminal His-tag (rAPRc$_{105-231}$-His) (Figure 15), no proteolytic activity was observed towards the same substrate. This result suggests that protease autoprocessing may indeed be accompanied by some conformational change that is not observed when the activation product is directly expressed in *E. coli*. Wan and co-workers have reported a similar result for HIV-1 PR by showing that a recombinant protein corresponding to the mature form of the protease (99 amino acids) with two additional amino-acids at the N-terminus (Met and Gly) displayed no proteolytic activity[247]. Based on this result, we have focused on the construct of the precursor form rAPRc$_{99-231}$-His for further analysis.

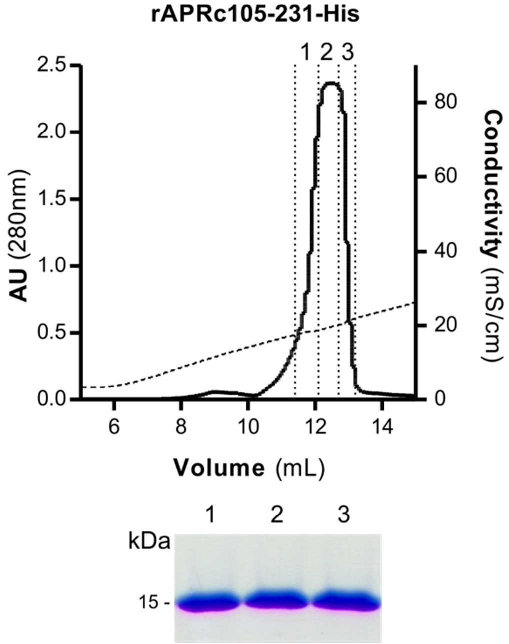


Figure 17. *Purification of rAPRc$_{105-231}$-His by cation exchange chromatography.* Recombinant rAPRc$_{105-231}$-His was first purified on a HisTrap 5 mL column as described for wild-type protease in legend of Figure 14. Protein elution on MonoS 5/50 GL column was carried out by a continuous salt gradient (0-1 M NaCl), at a flow rate of 0.75 mL/min and monitored by measuring the A_{280nm}. Approximately 4 µg of protein from fractions 1, 2 and 3 outlined by dotted lines in chromatogram were analyzed by SDS-PAGE stained with Coomassie blue. The molecular weight marker in kilodaltons (kDa) is shown on the left.

Dimerization studies on APRc

In view of the homodimeric nature of retropepsins, the oligomeric organization of both purified rAPRc$_{99-231}$-His as well as the derived activation product (APRc$_{105-231}$-His) was firstly evaluated by analytical size-exclusion chromatography (SEC). As shown in Figure 18, the results were always consistent with a preferential accumulation as monomers with only a much reduced amount of protein eluting as oligomers, as given by the detection of a shoulder (but not a defined peak) prior the elution of the monomers.

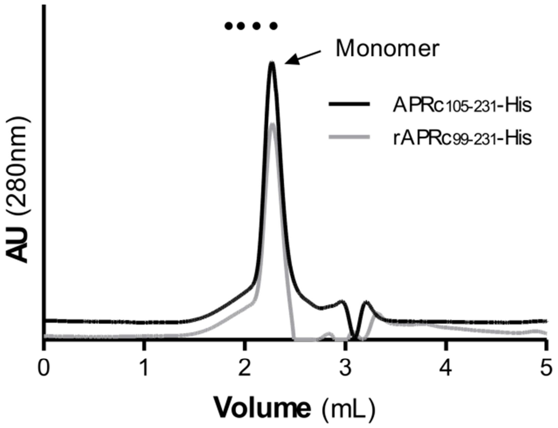


Figure 18. *Assessment of APRc oligomerization state*. The precursor rAPRc$_{99\text{-}231}$-His and activated APRc$_{105\text{-}231}$-His forms were analyzed by analytical SEC. The Superdex 200 5/150 GL was equilibrated in 20 mM phosphate buffer pH 7.5 containing 150 mM NaCl. The black dots refer to elution volumes of molecular mass markers used for calibration. From left to right: conalbumin (75 kDa), ovalbumin (43 kDa), carbonic anhydrase (29 kDa) and ribonuclease A (13.7 kDa).

For this reason, cross-linking studies were conducted to provide evidence for APRc dimer formation, as they are generally used to stabilize transient complexes and weak interactions. The use of cross-linking assays is based on the accepted premise that the interacting molecules must be in close proximity for a sufficient period of time and for a significant fraction of the population of molecules under study, in order to form a covalent bond between two cross-linkable residues[248]. In this study, reaction products and control samples from cross-linking assays performed with DSS, a homobifunctional amine-reactive cross-linker that covalently links Lys residues, were analyzed by immunoblotting (Figure 19A). As expected, the results revealed a significant amount of APRc associated as dimer, although monomeric and larger aggregate species were also visible. Strikingly, even though a conformational change is proposed for the conversion of rAPRc$_{99\text{-}231}$-His to the final form APRc$_{105\text{-}231}$-His, the two forms did not seem to markedly differ in their ability to self-associate in the dimeric state. To further confirm the molecular weight of the structures detected by Western-Blot, cross-linked and non-cross-linked protein samples were analyzed by analytical SEC, with the resulting chromatogram shown in Figure 19C. As expected, one major peak corresponded to the elution of the monomer on both samples, while the cross-linked sample exhibited an additional peak corresponding to the elution of the dimer with an apparent molecular weight of 30 kDa.

The use of cross-linkers with a long spacer harm usually yields higher efficiency of cross-linking since it is more likely that two reactive sites are within the distance range of the reagent[248]. However, this can also lead to lower specificity as the result of the cross-linking of molecules that are not in direct contact with each other but that can randomly collide under certain conditions (e.g., high protein concentration)[249]. Therefore, similar studies were also conducted using the cross-linker glutaraldehyde, which differs from DSS in length of connecting backbone (11.4 Å for DSS and 7 Å for glutaraldehyde) and reactive groups (Figure 19A-B). Despite reducing the chain length of the cross-linker spacer, the results were similar to those obtained with DSS further emphasizing the proximity of cross-linkable sites (Figure 19B). Hence, different primary amines might react with each cross-linker agent and/or a weak interaction between the two monomers might allow the required flexibility to accommodate both cross-linker agents.

Altogether, these results strongly indicate that APRc is mainly a monomer is solution, although there is a slight equilibrium between monomeric and dimeric states. Therefore, much like it has been described for other retropepsins from spumaretroviruses family[250,251], APRc appears to form weak transient dimers that might be only present under certain conditions, thereby consisting of a low fraction of the whole population of APRc oligomeric states.

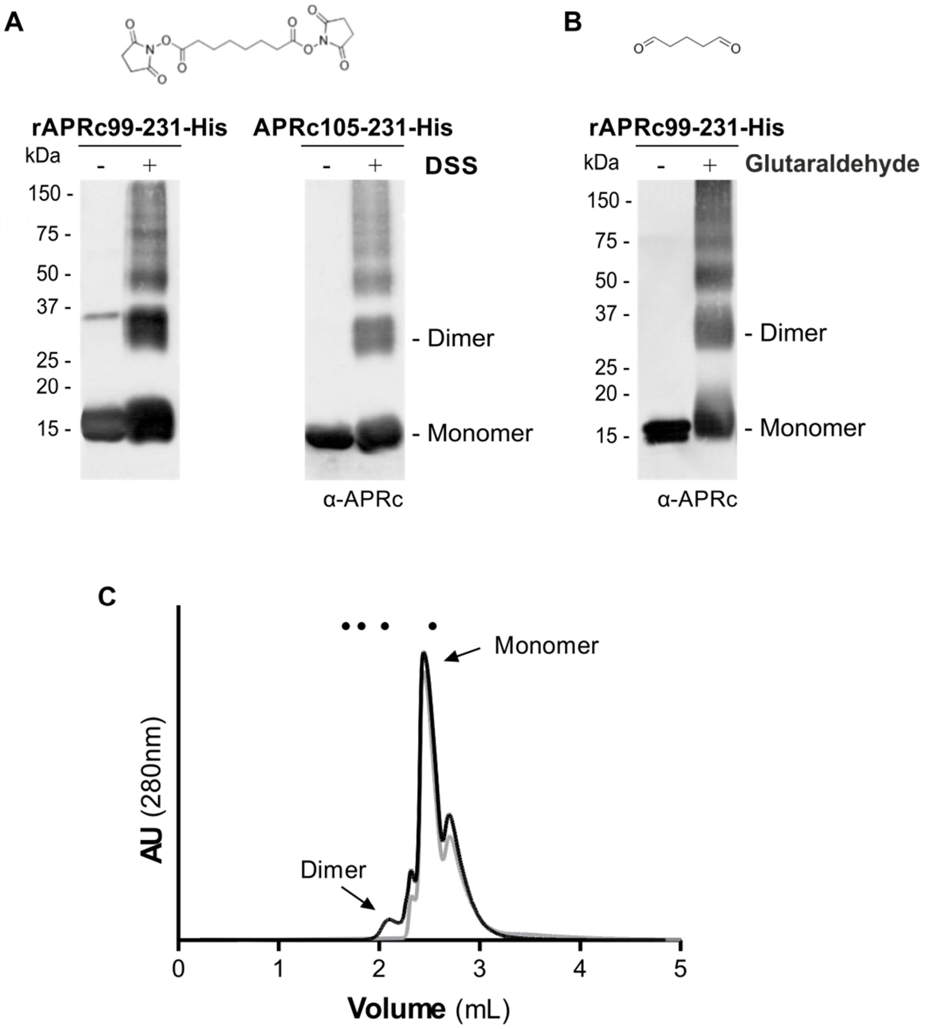


Figure 19. *Cross-linking studies of the last intermediate rAPRc$_{99-231}$-His and final form of activation (APRc$_{105-231}$-His).* (**A**) The quaternary configuration of rAPRc$_{99-231}$-His precursor and activated forms was assessed by incubating the protease with the cross-linker DSS. Both DSS treated and untreated protein samples were subjected to Western blot analysis with anti-APRc antibody. In the presence of the cross-linking agent, a significant proportion of the protein migrated as a dimer, although the monomeric forms and larger aggregates were also observed. (**B**) Despite the differences in the structure of cross-linking agent, similar results were obtained with glutaraldehyde as the cross-linking agent. (**C**) DSS treated and untreated rAPRc$_{99-231}$-His protein samples were applied to a Superdex 200 5/150 GL equilibrated in 20 mM phosphate buffer pH 7.5 containing 150 mM NaCl. The black dots refer to molecular weight of protein standards used for column calibration, from left to right: conalbumin (75 kDa), ovalbumin (43 kDa), carbonic anhydrase (29 kDa) and ribonuclease (13.7 kDa).

Enzymatic properties of active APRc

Based on the observed enzymatic activity upon conversion of the precursor form rAPRc$_{99-231}$-His, all characterization studies were focused exclusively on this derived activation product (for simplification APRc). The effect of pH was determined using the same fluorogenic substrate – [MCA]-Lys-Ala-Leu-Ile-Pro-Ser-Tyr-Lys-Trp-Ser-Lys-[DNP] - which mimics the identified auto-cleavage site between Ser_{104}-Tyr_{105} residues, in a range of pH values from pH 4 to pH 9. From this analysis an optimal activity at pH 6.0 was observed (Figure 20A), with no appreciable hydrolytic activity below pH 5.0. This higher optimal pH value is consistent with optimum pH values reported for other retropepsin-type proteases. Examples include the Walleye Dermal Sarcoma Virus PR which display an optimal activity at neutral pH for both the auto-cleavage and processing of Gag peptide substrate[252]. Other retropepsins such as HIV-1[253], EIAV and MLV (Murine leukemia virus) PRs have been also reported to have optimal activity in the range pH 4-6, which varies depending on the condition of the assays and type of substrate (e.g., protein versus peptide, or the presence of ionizable side chains such as Glu)[66].

When investigating the susceptibility of APRc to classical protease inhibitors (Figure 20B), this protease was shown to be insensitive to pepstatin, even though a slightly inhibitory effect was observed during autolytic processing. In contrast, APRc activity was strongly inhibited by EDTA retaining only 26% activity and a small inhibitory effect was also observed with Pefabloc. No substantial effect was observed after incubation with E-64 whereas incubation with Zn^{2+} (Figure 20B) slightly affected enzyme activity.

In order to provide additional evidence that APRc is indeed a retropepsin-like enzyme we analyzed the effect of different clinical inhibitors of HIV-1 PR (Figure 20C). Strikingly, incubation with indinavir resulted in a near complete inhibition of APRc, even when tested at a final concentration of 0.25 mM in the assay. Additionally, nelfinavir, saquinavir, amprenavir and atazanavir also had a remarkable inhibitory effect, ranging between approximately 30-50% of inhibition (Figure 20C). With the exception for amprenavir, when using higher concentrations than 0.25 mM of remaining inhibitors solubilized in DMSO (atazanavir, lopinavir, nelfinavir, ritonavir and saquinavir) aggregates were observed and, thus, these results were not included in this analysis. This inhibitory effect of specific HIV-1 PR inhibitors against a prokaryotic retropepsin-like enzyme has not been previously described.

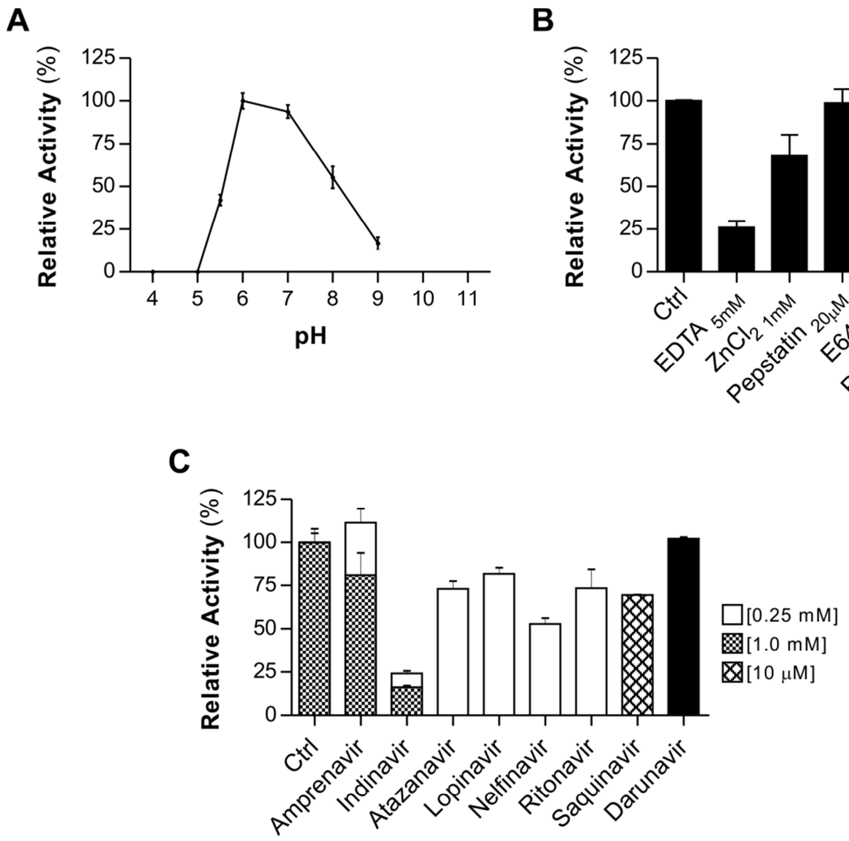


Figure 20. *pH and inhibition profile of APRc activation product.* The effect of pH, class-specific and HIV-1 PR specific inhibitors on the proteolytic activity of APRc activation product was evaluated using the synthetic fluorogenic substrate (MCA)-Lys-Ala-Leu-Ile-Pro-Ser-Tyr-Lys-Trp-Ser-Lys-(DNP). (**A**) Activity at different pH values. Activated APRc was incubated with the substrate at 37 °C in buffers ranging between pH 4 and pH 9 containing 100 mM NaCl (50 mM sodium acetate pH 4.0, 5.0, 5.5 and 6.0 and 50 mM Tris-HCl pH 7.0, 8.0 and 9.0), displaying an optimal activity at pH 6. (**B**) and (**C**) To test the effect different compounds, the protease was pre-incubated in the presence of each inhibitor for 10 minutes at room temperature in 50 mM sodium acetate pH 6.0 containing 100 mM NaCl before adding the substrate. APRc activation product was strongly inhibited by specific HIV-1 PR inhibitors, with the most prominent effect observed for indinavir. The rate of substrate hydrolysis (RFU/sec) was monitored for 3 hours and the relative activity normalized by setting the maximum activity at 100%. The error bars represent standard deviation of the mean.

2.4. Discussion

The intrinsic difficulty in working with obligate intracellular parasites such as rickettsiae greatly hampers the correlation of rickettsial gene products with their function. Therefore, valuable information on the nature of conserved genes as well as on the identification of new bacterial factors that may play a role in rickettsiae pathogenesis is mostly being provided by comparative genomics. Using this approach, we identified a gene encoding a putative membrane embedded aspartic protease with a retroviral-type signature, highly conserved in 55 *Rickettsia* genomes. Using the *R. conorii* gene homologue RC1339 as our working model we demonstrate that the gene product (APRc) displays a high degree of identity among *Rickettsia* spp., although no significant homology is observed when compared to other APs, except for the conservation of the motif around the catalytic aspartate as well as the hydrophobic-hydrophobic-glycine motif required for the formation of the psi loop. These features resemble the retroviral APs comprising family A2, which are characterized by being active only as symmetric dimers with a single active site, where each monomer contributes with one aspartate[27,41]. Despite the observed low overall sequence similarity with retropepsins, our results on the enzymatic characterization of RC1339/APRc soluble catalytic domain further revealed that this novel rickettsial enzyme indeed shares several properties with this family of APs, as discussed hereinafter.

Most viral retropepsins are strictly required for the processing of Gag and Gag-Pol polyproteins into mature structural and functional proteins (including themselves) and are, therefore, indispensable for viral maturation[91]. Because of this, retropepsin PRs are generally characterized by their inherent autolytic function. Interestingly, our results with APRc soluble catalytic domain fused to GST also demonstrated the ability of this protein to undergo a multi-step autocatalytic conversion *in vitro* into APRc$_{105\text{-}231}$ mature form, and this autolytic activity was again confirmed when the last intermediate of activation was produced in *E. coli*. Moreover, to investigate whether the lower molecular weight band observed for the purified sample of wild-type rAPRc$_{99\text{-}105}$-His do indeed correspond to the last product of autoproteolytic processing observed for rGST-APRC$_{87\text{-}231}$, three mutants of the last intermediate were generated by site directed mutagenesis. The mutation of the catalytic aspartate (rAPRc(D140A)$_{99\text{-}105}$-His) had the effect expected for a retropepsin-like enzyme, yielding an inactive enzyme with impaired autoprocessing. CD analysis has also confirmed the significant role of this residue for the structural stability of APRc, as previously reported for other retroviral proteases[49,244,245]. The other two mutants with altered internal cleavage recognition sequences to the last cleavage site between Ser104-Tyr105 residues, rAPRc(S104P)$_{99\text{-}105}$-His and rAPRc(Y105P)$_{99\text{-}105}$-His, respectively, while having a near-identical T_M (from CD analysis), exhibited very little tolerance of changes in this recognition sequence (only slight activation was observed for rAPRc(S104P)$_{99\text{-}231}$-His), thereby confirming that this is indeed the target for autoproteolytic processing.

The enzymatic activity assays performed during these auto-activation studies (either using oxidized insulin β chain or the fluorogenic peptide mimicking the final cleavage site between the Ser104-Tyr105 residues) clearly indicated that APRc activity appears to be dependent on the presence of the final activation product. These results suggest that the processing at the N terminus must be the determining step for the regulation of enzymatic activity, presumably through a conformational change occurring upon conversion from rAPRc$_{99\text{-}231}$-His to APRc$_{105\text{-}231}$-His form.

This is in line with what has been described for recombinant HIV-1 PR, where the increase in catalytic activity upon protease autolytic conversion has been correlated with a conformational rearrangement between the precursor/inactive vs. mature/active forms of the enzyme[91,254]. In the context of Gag-Pol precursor, the PR domain is flanked by the TFR and the RT enzyme at its N and C termini, respectively, where the TFR is thought to have a role similar to that observed at the N termini of zymogen forms of pepsin-like enzymes[246]. As represented in Figure 21, even though the full-length TFR/PR precursor appears to be monomeric, it undergoes maturation through intramolecular cleavage of a putative transient dimer[34,48,63,246]. For autoprocessing to occur, the dimer intermediate undergoes a conformational change, in which one of the two N-terminal strands occupies its active site. The dimer formation is facilitated by interface interactions of at least the active site and the C-terminal residues and possibly stabilized further by the interaction of the N-terminal TFR/PR cleavage site sequence with the active-site and flap residues[64,255]. Cleavage of the scissile peptide bond at the N-terminus of HIV-1 PR is the rate-determining step for the appearance of enzymatic activity towards Gag-Pol precursor. In accordance, and similar to what we have observed for APRc, the expression of a wild-type HIV-1 PR extended only with the initiator Met, lacking the original N-terminus of the protease consisting on the Phe-Pro bond, resulted on an inactive enzyme[247].

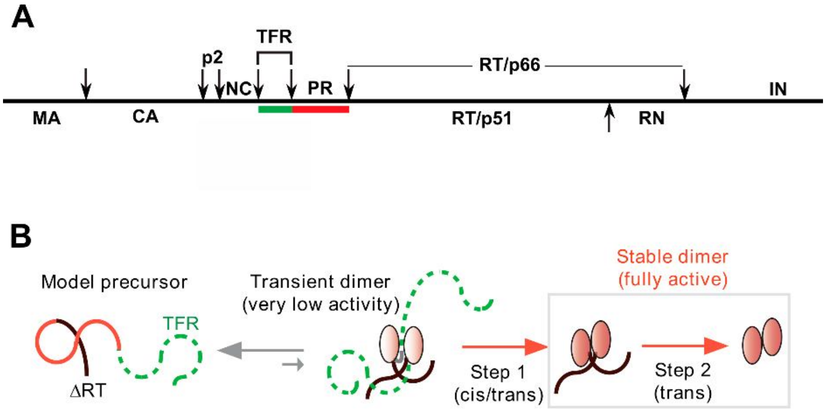


Figure 21. *Structural organization of the Gag-Pol polyprotein of HIV-1 and representation of its auto-processing mechanism.* (**A**) Individual domains are matrix (M); capsid (CA); nucleocapsid (NC); reverse transcriptase (RT); RNase H (RN); integrase (IN). Black arrows indicate specific sites of cleavage by PR. The green (TFR) and red (PR) bars denote the protease precursor (TFR-PR). (**B**) Proposed mechanism for the processing of a model precursor comprising the TFR, PR, and truncated ΔRT domains. Ovals indicate folded monomers in the transient precursor dimer, PR-ΔRT and mature PR. Transitioning ovals from light red to fully red depict appearance of catalytic activity and stable dimer formation of PR. Adapted from Louis et al.[65].

Despite all retroviral proteases have been previously shown to be only active as homodimers, the mechanisms governing the formation of dimer state are rather distinct among the two subfamilies of retroviruses: Orthoretrovirinae (e.g., HIV-1, RSV, etc.) and Spumavirinae (e.g., HFV and simian foamy virus). As discussed for HIV-1 PR, the presence of the N-terminal flanking TFR sequence leads to the formation of weak dimers with low PR activity. Once the N-terminal region is cleaved off, stable and active PR dimers are subsequently formed. In contrast, this type of regulation cannot take place with the foamy virus PRs as there is no Gag–Pol precursor and thus no N-terminal extension of the PR. Foamy viruses (FVs) express their Pol polyprotein from a separate Pol-specific transcript and, since only integrase domain is cleaved off, the mature protease harbors the reverse transcriptase at its C-terminus (PR-RT)[256]. As a result, it was reported that the PRs from simian foamy virus from macaques (SFVmac) have significant differences in the dimerization interface relative to most common orthoretroviral proteases counterparts (e.g., HIV-1 PR), even though PRs from both retroviruses families exhibit similar overall folds[250,251]. In fact, the existence of a predominant monomeric state of SFVmac PR-RT is supposed to inhibit PR activity before virus assembly in order to have a properly packaged viral unit[250]. To explore the possible influence of RT on dimerization, a separate PR domain (PRshort) of SFVmac was expressed in *E. coli*, and compared to the full-length PR-RT. Although only monomeric species could be detected when analyzed by SEC or analytical ultracentrifugation, both enzymes exhibited proteolytic activities in vitro at NaCl concentrations of 2–3 M. As SFVmac PR dimerization is required in order to form the catalytic center of this enzyme, recent investigations on monomer/dimer status of PRshort by paramagnetic relaxation enhancement analysis, have identified a transient homodimeric state which is characterized by an equilibrium between a lowly populated short-lived state in high-dynamic exchange with the ground state[251]. Given the largely monomeric state of PRshort and PR-RT under physiological conditions, it is likely that FV PRs requires additional viral and/or cellular factors for efficient dimerization *in vivo*. Attending to these data, APRc appears to share features with proteases of the two retroviral subfamilies: while it appears to display a similar type of regulatory activation mechanism reported for HIV-1 PR (requiring processing and conformational change prior achieving full enzymatic activity), this enzyme also revealed to be present in solution mainly under a monomeric state, although a slight amount of protein appears to be in dimeric form. Noteworthy, APRc also share with FV PRs the lack of the conserved Gly-Arg-Asn/Asp sequence motif present in most retroviral proteases. This motif has been described to have a significant role in dimerization, as the intra-monomer hydrogen bound between Arg87 and Asp29 (in HIV-1 PR) appears to influence the correct placing of the C-terminal β-strands (residues 96–99 in HIV-1 PR) at the interface of the two monomers[257]. Therefore, the absence of this important structural feature in APRc is most likely one of the key reasons why the dynamics of dimer formation and stability of this protease differs so much from the majority of retropepsins.

Further studies are, therefore, required to better understand the maturation and dimer formation of APRc precursor forms *in vitro* and how this is accomplished and controlled *in vivo*. In fact, as will be discussed in Chapter IV, we have shown that APRc accumulates in the outer membrane in *R. conorii* and *R. rickettsii* and, therefore, we cannot rule out that the presence of the transmembrane domain may have an important role in this maturation process *in vivo*. Actually, the presence of transmembrane domains in APRc introduces a new feature still poorly characterized in the context of retropepsin-like enzymes. A similar domain organization - putatively membrane embedded with a soluble catalytic domain - with variations in the number of predicted TMH has been also described for eukaryotic retropepsin-like proteases such as human and mouse SASPase[100,104] as well as for SpoIIGA from *Bacillus subtilis*[98]. SASPase is primarily expressed with a single transmembrane domain, as suggested by the immunodetection of unprocessed SASPase in the stratum corneum of psoriatic epidermis, but due to its auto-catalytic activity, only processed forms are found in the upper layers of normal skin[100,104]. In contrast, SpoIIGA is proposed to have five transmembrane segments in its N-terminal domain, which might function as a receptor for the SpoIIR signal[98]. Interestingly, by expanding the search to other gram-negative bacteria, Teixeira[258] identified several putative genes encoding for putative retropepsin-like enzymes not only in other α-proteobacteria but also in δ-, γ- and β-proteobacteria, all comprising putative transmembrane domains. Although a remarkable similarity in amino acid sequence was observed for these putative proteins in what concerns the hallmark catalytic sequence motifs - hydrophobic-hydrophobic-Asp-Thr/Ser-Gly-Ala and hydrophobic-hydrophobic-Gly - a clear difference was found regarding the number of putative transmembrane domains. When searching for proteins related to the retropepsin-like enzymes from *Rickettsia* spp., many other α-proteobacteria, such as *Polymorphum gilvum*, *Sinorhizobium fredii* and *Mesorhizobium amorphae*, were also identified as having genes coding for a putative soluble catalytic motif with three putative transmembrane helices (3TMH). Moreover, despite belonging to δ-proteobacteria, a putative retropepsin-like enzyme from *Desulfatibacillum alkenivorans* also shares the same 3TMH motif. Importantly, a BLAST search for retropepsin-like enzymes in γ-proteobacteria identified related proteins with a single transmembrane domain (1TMH) in many pathogenic bacteria including *Pseudomonas* ae*ruginosa, Pseudomonas putida* and *Legionella pneumophilla*. Besides the γ-proteobacteria matches, this search also retrieved proteins with 1TMH in α-proteobacteria such as PerP (*Caulobacter crescentus*), in β-proteobacteria (*Thiobacillus denitrificans* and *Ralstonia* sp. 5_7_47FAA) and in δ-proteobacteria (*Syntrophus aciditrophicus* and *Desulfobacter postgatei*)[258]. From this analysis it is obvious that the presence of transmembrane domains is a common feature among putative prokaryotic retropepsin-like proteins, definitely urging for future studies on the functional relevance of this structural organization.

Another interesting observation was that APRc autolytic activity, as well as cleavage of the fluorogenic substrate, occurred at a pH optimum of 6.0. This is again in good agreement with the optimal pH of other retropepsin-like[252,253] enzymes as well as of the pepsin-like renin[259,260] and, actually, it is consistent with the presence in APRc of an alanine residue downstream from the catalytic motif (Asp-Thr-Gly-<u>Ala</u>), instead of the common Thr residue found in most pepsin-like APs. This substitution affects the acidity of the active site residue Asp by preventing a hydrogen bond forming[41]. Unexpectedly, we observed a drastic inhibitory effect of EDTA on both APRc maturation and hydrolysis of the fluorogenic substrate, suggesting that this protease may depend on a metal ion for folding and/or activity. A similar effect has not been reported for other retropepsins and no homology to a metalloprotease consensus motif was identified in APRc that could justify this inhibition.

It is also important to note that a distinct pattern of inhibition of pepstatin and indinavir was observed between the maturation process and the enzymatic activity towards other substrates (e.g., PepRick14). These two AP inhibitors exhibit remarkable differences in structural composition: pepstatin is a hexa-peptide with the sequence: Iva-Val-Val-Sta-Ala-Sta, where Sta is the unusual amino acid statine, whereas indinavir is a structural analogue of the HIV-1 PR Phe-Pro cleavage site. Therefore, the aforementioned suggested conformational changes on the active pocket site of APRc upon the conversion of the last intermediate to the mature form, may once again justify the ability of APRc to accommodate one or another inhibitor. Interestingly, whereas these two inhibitors have been reported to inhibit different retropepsins (e.g., HIV-1 and XMRV PRs[65,246,261]) either on both maturation and enzymatic activity, a contrasting effect was also observed for the auto-activation of SASPase[100]. In fact, even under the presence of 1 mM pepstatin, no inhibitory effect was observed for SASPase auto-activation while the inhibition by indinavir was suggested to be responsible for the skin side effects observed in patients during therapy for AIDS treatment[100].

Given the remarkable inhibitory effect of indinavir on APRc activity, important insights regarding the contribution of APRc for rickettsial infection mechanisms may be obtained in the future by evaluating the impact of indinavir, and other commercially available HIV-1 PR inhibitors, on the ability of different rickettsial strains to adhere and invade to the host cells. Such information would greatly assist in the assessment of the potential of APRc as a candidate therapeutic target for the treatment of rickettsioses, and consequently in the development of specific drugs directed to APRc inhibition.

Further structural studies on APRc will be required to help understanding the molecular mechanisms of activation, the different susceptibilities to inhibitors as well as to provide additional insights into structure-function relationships of this novel aspartic protease. Nevertheless, the observed autolytic activity, optimal activity at mildly acidic pH and the observed inhibitory effect of specific HIV-1 PR inhibitors, clearly strengthen the striking resemblance between the enzymatic properties of APRc and those of viral retropepsins. Moreover, the results described here provide experimental substantiation that RC1339/APRc is a novel functional retropepsin-like enzyme expressed in a gram negative intracellular bacterium, contributing to the analysis of the evolutionary relationships between the two types of APs that will be further discussed in Chapter V.

Chapter III. APRc specificity profiling by PICS

Chapter III. APRc specificity profiling by PICS

3.1. Introduction

One of the key factors in maintaining the fidelity of most biological features is the substrate specificity of proteases, i.e., their ability to discriminate among many potential substrates. The substrate specificity of a protease can be controlled on many levels in a biological context, including the spatial and temporal localization of the protease and their potential substrates, the post-translational modifications of proteases or the requirement of cofactors or adaptor proteins, but it is mostly determined by the organization and composition of the substrate-binding subsites in the catalytic pocket[262]. Although the primary specificity of a protease is determined by the amino acid residue that is accepted in its S1 (or S1') position, the cleavage site selectivity of proteases largely relies on the recognition of substrate residues spanning the scissile peptide bond. Therefore, the conformation and flexibility of the peptide backbone of the substrate protein also significantly affects the efficiency of the peptide bond cleavage[263]. These and many other biochemical properties of proteases have been unveiled by the increasing number of studies on the role of these enzymes and associated functional mechanisms in biological processes, which include mapping the specificity of proteases and the identification of their substrates and inhibitors. The recognized importance of these studies led to the definition of the term *Degradomics* by Lopez-Otin and Overall in 2002[264]: "*All genomic and proteomic investigations and techniques regarding the genetic, structural and functional identification and characterization of proteases, and their substrates and inhibitors, that are present in an organism".*

Protease specificity profiling, in particular, is a key step in the biochemical characterization of proteases, and also provides valuable information on the active-site structure to help in the design of specific peptide substrates (useful for assay development) and of small molecule protease inhibitors. Until recently, specificity studies of proteases were typically based on phage display[265] or peptide library approaches[266,267]. However, mass spectrometry-based proteomic strategies are now becoming widely adopted approaches for protease cleavage sites and activity profiling[268]. Their main advantages over other methods are their high sensitivity and selectivity, their ability to identify and characterize the primary structure of the cleavage products and, depending on the method, their ability to perform quantitative analysis[269]. PICS (Proteomic Identification of protease Cleavage Site specificity) is a recent peptide-centric approach for the easy mapping of endoprotease subsite preferences that allows the identification of both prime- and non-prime side specificity in the same experiment[268,270,271]. This is a unique technique in the sense that it uses relevant and complex proteome-derived oligopeptide libraries and thus harnesses natural sequence diversity. The schematic diagram from Figure 22 illustrates the main steps of PICS methodology[271]. PICS libraries can be generated from any proteome source with known genome or proteome sequence (e.g., human or *E. coli* cells), and thus represent biological sequence diversity. A specific endoprotease, such as Trypsin, GluC or Chymotrypsin, is used to digest the cellular proteomes and following protease inactivation, thiols and primary amines (N-terminal α-amines and Lys ε-amines) are chemically protected. The proteome-derived peptide library is then cleaned up and purified. Such libraries are now ready to be used as substrate for the PICS assay with a test protease. After incubation, in contrast with the non-cleaved peptides in the library that possess blocked primary amines, each prime-side (P') cleavage fragment has a reactive primary amine on their neo-N termini, which is biotinylated by NHS chemistry for affinity isolation and sequence identification by liquid chromatography–tandem mass spectrometry (LC-MS/MS).

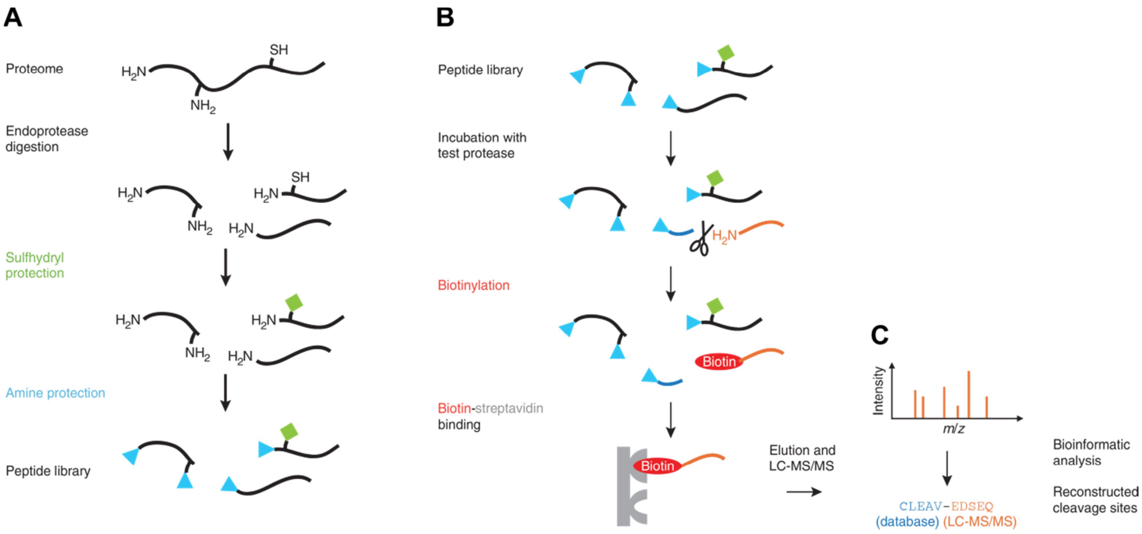


Figure 22. *Schematic representation of PICS methodology.* (**A**) Libraries are generated from cellular proteomes through digestion into oligopeptides by specific endoproteases. Thiols and primary amines are then chemically blocked and the library is purified. (**B**) PICS peptide libraries are incubated with a test protease and the newly formed prime-side cleavage products, which possess free N termini, are then tagged with a cleavable biotin, allowing specific isolation with immobilized streptavidin. (**C**) After elution, prime-side cleavage products are identified by LC-MS/MS while the corresponding non-prime sequences are determined using bioinformatics. Adapted from Schilling and colleagues[271].

Since it relies on database-searchable peptide libraries, the corresponding non-prime sequences are derived bioinformatically through a web-based service, allowing the simultaneous determination of the amino acid residues spanning both sides of the scissile bond: the prime-side (P') and non-prime side (P)[271,272].

One major limitation of this approach is that it has been designed for a qualitative determination of cleavage sites specificities but does not allow analyzing quantitative aspects of proteolytic reactions that would allow the comparison of cleavage site specificity of the same protease in different conditions (e.g., different pH or temperature). Also, this approach cannot be used for exopeptidases, which requires the use of available synthetic peptide libraries covering all the amino acids combinations that can be used. Finally, although PICS libraries contain all natural amino acids, the preparation of peptide libraries implies that cysteines and lysines must be blocked, thereby conferring them different properties that must be taken into account when analyzing the output results. Regardless these limitations, the large number of cleavage sequences identified in each PICS experiment, allows not only for the identification of protease cleavage sites but also for the analysis of potential subsite cooperativity, as already demonstrated for HIV-1 PR and for other classes of proteases[270].

The results presented in the present chapter extend and validate the applicability of PICS to APs (HIV-1 PR is the only AP with specificity profile determined by PICS approach) by providing the identification of amino acid preferences of APRc. The comprehensive analysis of APRc specificity profile and further comparison with the specificity of other APs, clearly corroborates its identity as a new enzyme belonging to the retropepsin family.

3.2. Materials and Methods

PICS methodology

Protease specificity profiling with proteome derived peptide libraries consisted of library preparation and PICS assay as described elsewhere[270] with minor changes.

Library generation

PICS libraries were generated with THP1 cells derived from the peripheral blood of a 1-year-old male with acute monocytic leukemia. The cells were grown in suspension in three 175-cm^2 cell culture flasks to a density of 2-4 x 10^8 cells/mL to yield a total mass of approximately 0.5 g cells. Cells were centrifuged at 400*g*, 4 ˚C for 5 min and the supernatant decanted. The cell pellet was washed with PBS and the centrifugation repeated.

Once determined the weight of the cell pellet, it was re-suspended in 5 volumes (assuming that 1 mg cell pellet ≈1 mL volume) of hypotonic lysis buffer (100 mM HEPES pH 7.5, 1 mM PMSF, 10 mM EDTA, 0.1% SDS) containing the proper amount of Thermo Scientific Halt Protease Inhibitor Single-Use Cocktail EDTA-free solution according to the manufacturer's instructions. To promote cell lysis, a repeated aspiration (15 or more times) of the cell suspension with a 22 or 27-gauge needle was performed. The cell lysate was centrifuged at 20000*g* for 20 min and the supernatant collected to a new 50 mL falcon tube and adjusted to 100 mM HEPES pH 7.5. Protein concentration was determined by measuring the Abs_{280nm}.

For the first round of sulfhydryl alkylation, the sample was incubated with 5 mM DTT for 1 h at 25 ˚C. The sample was then incubated with 20 mM iodoacetamide for 1 h at 25 ˚C in the dark and with 5 mM DTT at 25 ˚C for 15 min to quench unreacted iodoacetamide.

Protein precipitation was done by the addition of chloroform/methanol according to Wessel and Flugge[273]. The procedure begun with the addition of 4x sample volume of cold methanol followed by the addition of 2x original sample volume of chloroform and 2x original sample volume of 50 mM Tris-HCl containing 150 mM NaCl. Between each addition, the sample was mixed by vortexing and then centrifuged for 1 min at 9000*g*. Upper and down phases were decanted while the protein pellet formed at the interphase remained attached to the wall of the falcon tube. The pellet was then washed twice with minute amounts (2x sample volume) of –20 ˚C cold methanol and left to air-dry for 30-40 min in the hood.

Subsequently, the pellet was overlaid with ice-cold 20 mM NaOH with sufficient volume to reach an assumed 2.0 mg/mL protein concentration on the basis of the total protein amount previously determined. To neutralize sample pH, 200 mM HEPES pH 7.5 was added before determining the protein concentration and total protein amount using the Bradford assay method.

Proteomes were digested either with Trypsin and GluC to generate tryptic or GluC libraries using a protease-to-proteome ratio of 1:100 (wt/wt) for Trypsin and 1:50 for GluC and incubated at 37 ˚C for 16h. To guarantee a complete digestion, equal amounts of enzyme were added for 2 h at 37 ˚C yielding a final ratio of 1:50 for Trypsin and 1:25 for GluC. To assess the completion of the proteome digest, a small aliquot was analyzed by SDS-PAGE, where no major protein bands above 10 kDa were observed after staining. Given the complete digestion, 1 mM PMSF was added to abolish activity of the digestion protease and 1 M guanidine hydrochloride to disrupt the formation of small aggregates. Bigger aggregates were pelleted by centrifugation at 20000*g* at 4 ˚C for 10 min.

The second round of sulfhydryl alkylation consisted on the incubation of digested proteome with 5 mM DTT at 37 ˚C for 1 h followed by the incubation with 40 mM iodoacetamide at 37 ˚C for 1.5 h in the dark. Finally, the sample was incubated with 15 mM DTT at 37 ˚C for 10 min.

For the free amine dimethylation, the peptides were incubated with 30 mM formaldehyde and 30 mM sodium cyanoborohydride ('ALD coupling solution') at 25 ˚C for 2 h. An additional 30 mM formaldehyde and 30 mM sodium cyanoborohydride were added and incubated at 25 ˚C for 16 h. The excess of free sodium cyanoborohydride was captured by incubating the sample with 100 mM glycine at 25 ˚C for 30 min.

The sample was acidified (pH < 2.5) to 2% (vol/vol) TFA purified by C18 solid-phase on a GE HealthCare RESOURCE™ RPC 3 mL column extraction according to manufacturer's instructions, with the exception of eluting with 80% (vol/vol) acetonitrile without added TFA. After elution, peptide concentration was determined by the BCA method.

The acetonitrile was then removed from solid-phase extraction eluate by vacuum evaporation. Throughout this step, the sample was refilled with water to approximately half of the original volume for four times. The sample was finally evaporated to achieve a calculated peptide concentration of 2.0 mg/mL. To redissolve peptides, the sample was incubated in an ultrasonication bath for up to 3 h followed by centrifugation at 20000*g* for 10 min.

Final peptide concentration was determined by the BCA method and peptide libraries were stored in aliquots of 200 µg at –80 ˚C. To confirm purity and integrity of the peptide library, 10 µg of the library was analyzed by ESI-MS/MS.

Test protease assay

One aliquot of peptide library was thawed and the appropriate buffer conditions for APRc incubation adjusted to have 50 mM sodium acetate buffer pH 6 containing 150 mM NaCl and a final library concentration of 1 mg/mL. Activated form of APRc (APRc$_{105-231}$-His) obtained as described in Chapter II, was incubated in a protease library ratio of 1:50 (wt/wt) for 16 h at 37 °C. To terminate the reaction, APRc was heat inactivated (90 °C, 5 min).

For the biotinylation of neo-amino termini, the pH was adjusted to 7.5 by adding 100 mM HEPES pH 7.4 and incubated with fresh 0.5 mM sulfosuccinimidyl 2-(biotinamido)-ethyl-1,3-dithiopropionate (sulfo-NHS-SS-biotin) at 25 °C for 2 h.

Biotinylated prime-side cleavage products were captured onto high-capacity streptavidin-sepharose (GE Healthcare) slurry previously equilibrated in 50 mM HEPES pH 7.4, 150 mM NaCl. Resin slurry volume was 1.5-times the volume of the PICS assay, providing sufficient binding capacity for bound and unbound biotin in PICS assay without further cleanup. Incubation was left at 4 °C for 16 h. The slurry was then washed by repeated (10x) centrifugation (200*g*, 30 seconds) and resuspension steps with 1 mL of 50 mM HEPES pH 7.4, 150 mM NaCl. The slurry was transferred to a spin column with a filter of approximately 10 µm pore size and the column was placed in a 2 mL reaction tube. The tube was centrifuged at 200*g* for 30 seconds and the flow-through re-applied to resin for a second centrifugation. After discarding the flow-through, the slurry was washed by repeated (10x) centrifugation at 150*g* for 30 seconds with 400 µl of washing buffer (50 mM HEPES pH 7.4, 150 mM NaCl). To elute the peptides, the slurry was gently resuspended in 50 mM HEPES pH 7.5 with 20 mM DTT, incubated for 1 h at 25 °C and finally centrifuged at 500*g* for 1 min. This step was repeated and the two elution fractions pooled together.

The eluted sample was further acidified with 2% formic acid (pH < 2.5) and loaded onto a reversed-phase C18 cartridge (100 cc) previously equilibrated with 1 mL 0.5% formic acid and centrifuged at 700*g* for 1 min. The column was washed with the same buffer by loading 10 times 1 mL. Finally, peptides were eluted with 1 mL 70% acetonitrile plus 0.5% formic acid and vacuum-evaporated to near-dryness (10-20 µl).

LC-MS/MS and data analysis

The proteomic identification of carboxy-peptide cleavage products by LC-MS/MS analysis was carried out using an nano-LC system (Thermo) with a 2 cm-long trap column (100 µm inner diameter, packed with 5 µm-diameter Aqua C-18 beads (Phenomenex)) and a 20 cm-long analytical column (50 µm inner diameter, packed with 3 µm-diameter Reprosil-Pur C-18-AQ beads (Dr. Maisch, Ammerbuch, Germany)) connected to a LTQ-Orbitrap XL hybrid mass spectrometer (Thermo) operated by the UBC Centre for High Throughput Biology), and on HALO^{TM} C18 column (Eksigent) connected to a LC-MS/MS TripleTOF 5600 (AB SCIEX), operated by the Center for Neuroscience and Cell Biology Proteomics Unit. For the QSTAR Pulsar instrument, samples were resuspended in 2% (vol/vol) acetonitrile, 0.1% formic acid and loaded onto a column packed with PepMap C18 resin (Dionex). Peptides were eluted using a 5-40% gradient of organic phase (buffer B) over 90 min. Buffer A was 2% acetonitrile and 0.1% formic acid, whereas buffer B was 85% acetonitrile and 0.1% formic acid. For the TripleTOF 5600 instrument, samples were resuspended in 0.1% formic acid and loaded onto a column packed with HALO^{TM} C18 column (Eksigent). Peptides were eluted using a 2–40% gradient of organic phase over 120 min. Buffer A was 0.1% formic acid and buffer B was 98% acetonitrile, 0.1% formic acid.

MS data was acquired automatically using Analyst QS software (Applied Biosystems) for information-dependent acquisition based on a 1 s MS survey scan (mass ranges listed below) followed by up to 3 (QSTAR Pulsar) or 2 (TripleTOF) MS/MS scans of 3 s each. Nitrogen was used as the collision gas and the ionization tip voltage was 22,000 V (QSTAR Pulsar) or 25,000 V (TripleTOF).

The identification of prime-side sequences from LC-MS/MS data was done with the spectrum-to-sequence database search programs Mascot and X!Tandem[274] in conjunction with PeptideProphet[275] at a 95% confidence level. Search parameters were set to identify static modifications as carboxyamidomethylation of cysteine residues (+ 57.02 Da), dimethylation of lysines (+ 28.03 Da) and thioacylation of peptide N termini (+ 88.00 Da). Mass tolerance was set to 10 ppm for the parent ion and 0.6 Da for fragment ions. Semi-style cleavage searches were applied with no constraints for the orientation of the specific terminus. Tryptic specificity was defined to cleavage C-terminal to Lys or Arg and GluC specificity was defined to cleavage C-terminal to Glu or Asp. Up to three missed cleavages were allowed for the library-generating enzyme.

Nonprime-side sequences of the original peptidic substrates were inferred by the web-based integrated series of data handling scripts termed WebPICS[271]. This web service automatically retrieved non-prime cleavage sequences, rendered the list of cleavage sites nonredundant, provided heatmap-style graphical and tabular representation of subsite preferences, and screened the cleavage sites for potential subsite cooperativity. The occurrences of amino acids relative to natural abundance retrieved by WebPICS were log-transformed in order to represent the under- and over-represented amino acids for each position with the same amplitude in opposite directions. Final heatmaps were constructed using MeV software[276].

The sequence logos representation of APRc subsite specificities encompassing the cleavage site were further obtained with IceLogo tool available at http://icelogo.googlecode.com/. Cleavage site sequences from P4 to P4' created by WebPICS were filtered to exclude redundant peptides and analyzed by the IceLogo algorithm using the proteome background (Swiss-Prot) of *Homo sapiens* as the reference set. Fold change and percentage difference graphical representations were generated with a *p*-value of 5%.

Heatmaps generation from *MEROPS* database

To provide a better means to compare the specificity profile of APRc with other APs, heatmap-style graphical representations of subsite preferences of selected APs from family A1 and A2 were constructed based on the annotations on *MEROPS* database[1]. The number of occurrences of each amino acid in each position around the scissile bond (from P4 to P4') were obtained from the specificity matrix of pepsin A (417 cleavages), cathepsin D (897 cleavages), BACE1 (24 cleavages), Mason-Pfizer monkey virus PR (18 cleavages), HIV-2 PR (30 cleavages) and feline immunodeficiency virus PR (28 cleavages). These values were then log-transformed in order to represent the under- and over-represented amino acids for each position with the same amplitude in opposite directions and the final heatmaps generated using MeV software[276].

3.3. Results

APRc specificity profile

To further biochemically characterize this novel rickettsial protease we determined its specificity profile by making use of the high-throughput peptide centric approach PICS[270,271]. In this work, active APRc was incubated with PICS libraries generated by digestion of total human THP1 cell proteins by either Trypsin or GluC. These PICS experiments resulted in the identification of 830 and 327 C-terminal cleavage products from tryptic and GluC libraries, respectively (Supplementary Table 1 and 2). The corresponding N-terminal cleavage products and complete cleavage sites were obtained and summarized using the WebPICS tool[271]. The PICS-based APRc specificity profiles are shown in Figure 23, and a good agreement was observed between the two complementary peptide libraries. APRc displays only moderate specificity and accepts multiple amino acids at each position. At P1, directly preceding the scissile bond, APRc showed a preference for large hydrophobic residues such as Phe, Tyr, Met, Leu, and carboxyamidomethylated Cys (modified during library preparation). In addition, APRc also preferred the neutral amino acids Thr and Asn at this site. A similar preference was observed for P1', although this further included small amino acids Ala, Ser, and Gly as well as Asp. Notably, cleavage sites were almost devoid of Pro at P1 and P1'.

Furthermore, PICS revealed distinct preferences for selected amino acids at other positions, likely reflecting structural constraints imposed by the substrate recognition and binding to the pocket site. In P2, APRc preferences include Val, Ile, Pro, and Thr, whereas a predominant preference for small and branched aliphatic amino acids Ala, Val, and Ile was observed at P2'. More distant from the cleavage site, small preferences for Val and Ile at P3, for Ala and Gly in P3', and a strong preference for Leu or Ile at P4' were observed. Interestingly, basic and acidic residues were significantly underrepresented throughout.

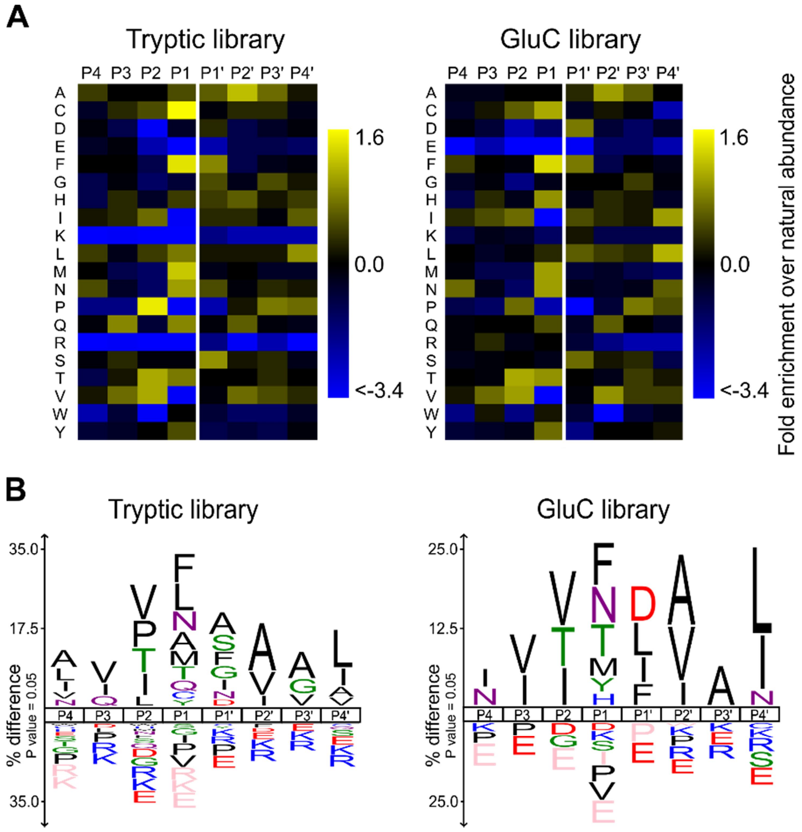


Figure 23. *APRc specificity profile determined by PICS methodology.* APRc specificity profile represented by Heatmaps and IceLogos, reveals similar amino acid preferences to both retropepsin and pepsin-like enzymes. Results are from tryptic and GluC peptide libraries derived from a *Homo sapiens* proteome (THP1 cells) incubated with activated APRc at a ratio of 1:50 (enzyme/library). The analytical strategy applied was similar to that described in [271]. PICS libraries were analyzed by multiple sequence alignments and applying correction for natural amino acid abundance. For each class of PICS library, the average amino acid occurrences in P4–P4' were calculated from three experiments and are either shown in the form of (**A**) a two-dimensional heat map of log(2) transformed values of fold-enrichment over natural abundance of amino acids and (**B**) % difference IceLogos. Both tryptic and GluC display consistency between them. In IceLogos representation, horizontal axis represents the amino acid position and vertical axis denotes the over-representation of amino acid occurrence compared with the Swiss-Prot *Homo sapiens* protein database. Cysteines are carboxyamidomethylated and lysines are dimethylated.

The large number of APRc cleavage sites identified from the tryptic PICS library further allowed investigation of subsite cooperativity. When comparing two of the strongest cleavage site determinants, Pro at P2 and Phe at P1, we observed apparent mutual exclusion (Figure 24). Of the 103 unique cleavage sites that contained Pro in P2, only 4 had Phe in P1 (3.7% compared to 10.5% occurrence for all identified cleavage sites), which was compensated by more frequent occurrence of P1 Met (10.3% compared to 5.8% total occurrence) and P1 Asn (14% compared to 8.3% total occurrence) (Figure 24A). Correspondingly, peptides with Phe in P1 yielded 4.6% P2 Pro (compared with 12.9% total occurrence) (Figure 24B), whereas peptides with Met or Asn in P1 yielded 22.9% or 21.7% P2 Pro, respectively (Figure 24C-D). A similar trend was observed in identified cleavage sites from GluC libraries, indicating subsite cooperativity between P2 and P1.

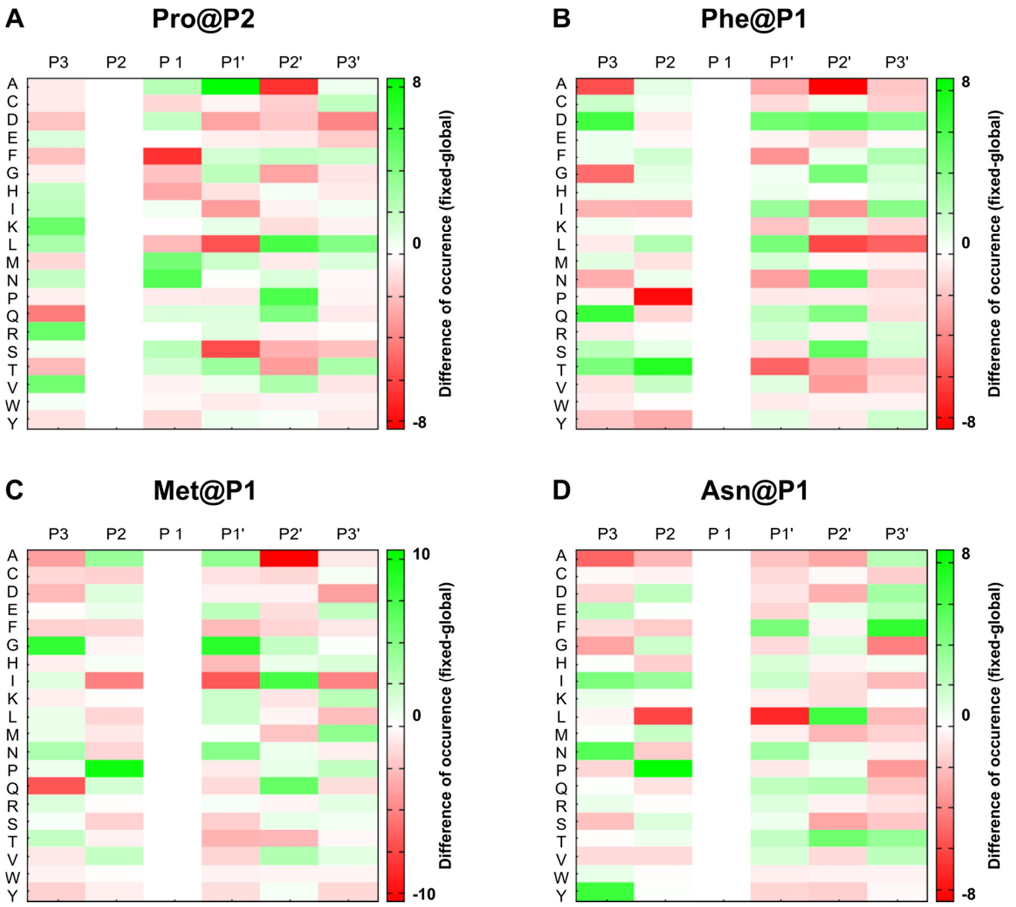


Figure 24. *Subsite cooperativity of APRc between P2 and P1.* The statistical analysis of subsite cooperativity of APRc was provided by WebPICS tool[277]. (**A**) When analyzing only cleavage sites with Pro in the P2 position (103 events), the number of cleavage sites with Phe in P1 dropped from 10.5% to 3.7%. (**B**) Consistently, 8.3% instead of 14% of cleavage sites revealed Pro in the P2 position in an analysis of cleavage sites with fixed Phe in P1. This negative frequency change indicates that Pro and Phe in these positions are not cooperative. (**C**) In contrast, the frequency of Pro in P2 rises to 22.9% with Met in P1, (**D**) and the same pattern is observed with Asn fixed in P1 position (21.7%). Therefore, these analyses unveil subsite cooperativity between Met and Asn in P1 and Pro in P2 position.

All together, these results clearly show that although displaying a unique profile, APRc shares some specificity requirements with retropepsins as well as with pepsin-like enzymes (particularly BACE1), further supporting APRc has being a member of the AP family.

3.4. Discussion

In view of the many advantages of mass spectrometry-based proteomic methods for mapping protease specificity, in this work we determined both prime and nonprime side specificity of APRc using PICS[270], providing additional evidence on the nature of this newly identified retropepsin-like enzyme. To date, HIV-1 PR is the only AP for which a PICS analysis has been reported[270]. However, several other studies have been carried out for many different APs on specificity towards individual substrates providing a collection of cleavage patterns for these enzymes. *MEROPS* database[1] compile, at least partially, the known cleavage sites in proteins and peptides (physiological and non-physiological) as well as in synthetic substrates collected from experimental data. *MEROPS* substrate cleavage collection shows a frequency matrix for the residues accepted in binding pockets P4 to P4', when 10 or more substrates are known. From this data, we have generated similar heatmaps as those obtained with PICS methodology for APRc (Figure 25), in order to compare the specificity profile of APRc with that of other APs. It is important to be aware, however, that the majority of substrates included in this collection have been identified from synthetic library peptides, some of which generated with fixed preferential positions. Therefore, the results are not fully comparable with APRc heatmaps but provide a glimpse on the major specificity determinants for each protease.

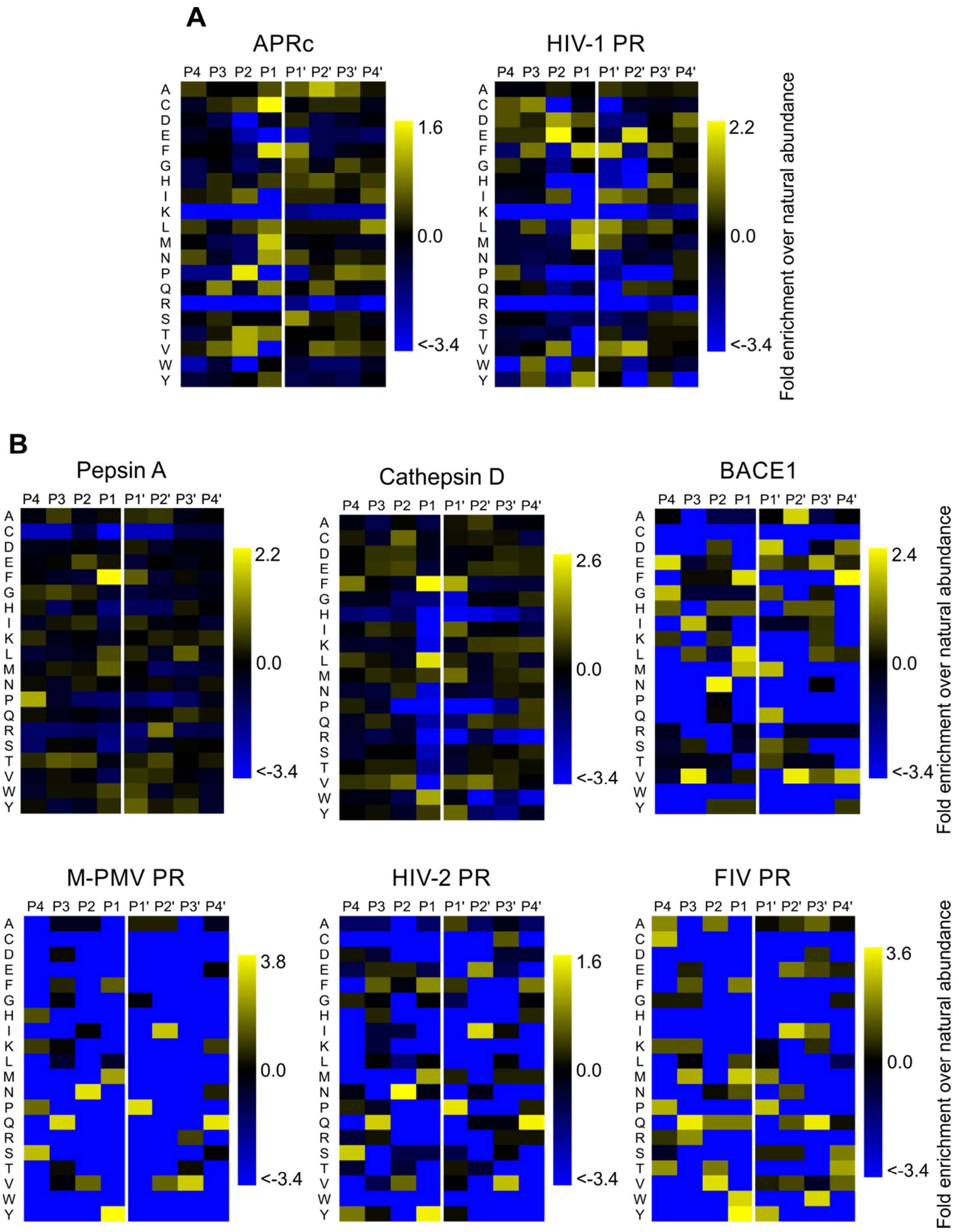


Figure 25. *Comparative analysis of APRc specificity profile with reported specificity of other aspartic proteases.* (**A**) The specificity profile of APRc derived from the digestion of human tryptic peptide libraries is compared with the corresponding specificity profile of human immunodeficiency virus 1 protease (HIV-1 PR) using PICS methodology, as described elsewhere[270]. Similarities between APRc and HIV-1 PR specificity profile were mostly confined to P1 and P1' position, with a common preference for hydrophobic amino acids such as Phe, Leu and Met. (**B**) Heatmap representations of other APs specificity profile were generated based on the annotations on *MEROPS* database of the corresponding substrate specificity: pepsin A, cathepsin D, BACE1, Mason-Pfizer monkey virus protease (M-PMV PR), human immunodeficiency virus 1 protease (HIV-2 PR) and feline immunodeficiency virus protease (FIV PR). Comparison of APRc specificity with that of other retropepsin and pepsin-like enzymes reveals a higher degree of specificity similarity with pepsin-like enzymes.

Even though the cleavage apparatus and the mechanism are nearly identical to all APs, the substrate specificity and the enzymatic sites accomplishing this specificity are vastly different. The comparative analysis of APs substrate specificity with our PICS results confirmed, nevertheless, common preferences between APRc and both retropepsin and pepsin-type APs. The amino acid preference of APRc for P1 position is in good agreement with the canonical specificity of APs for large hydrophobic amino acids, such as Phe, Met, carboxyamidomethylated Cys (which results from the modification during generation of peptide libraries), or Leu. Despite the observed lower selectivity, a similar trend for accommodating hydrophobic amino acids is also observed in P1'.

As observed in both tryptic and GluC libraries, APRc appears to display broader specificity for P1 and P1', while a more constrained amino acid preference is observed for P3, P2, and P2' positions. This observation may account for an important role on substrate recognition and binding to the active pocket site and may ultimately influence hydrolytic efficiency. Strikingly, a high degree of similarity is found with more specialized pepsin-like enzymes such as BACE1 for P3 (with a preference for Val and Ile) and P2' (Ala and Val) positions, as well as with cathepsin D (also for P2'). Interestingly, APRc also displays unique amino acid preferences such as Pro at P2 (although the preference observed for Val and Thr in this position has also been described for feline immunodeficiency virus PR[1,278]), and Leu and Ile in P4' position. Moreover, our results also suggest a cooperative effect between P2 and P1 positions by revealing that a P2 Pro co-occurs more frequently with P1 Met or Asn residues and that Pro is not favored at this position when P1 is occupied by Phe.

When compared with the two major types of cleavage sites proposed for HIV-1 PR and other retropepsins (type 1 and type 2), APRc specificity profile suggests a preference for type 2-like substrates with hydrophobic amino acids in P1 and P1', whereas type 1-like substrates with the typical combination of Tyr/Phe-Pro at P1-P1' appear disfavored[28,91]. As previously mentioned, comprehensive specificity studies as well as HIV-1 PR-inhibitor crystal structures have shown that the specificity of HIV-1 PR is strongly dependent on the substrate sequence context, which provides an explanation for the lack of a consensus sequences[66,81,82]. Likewise, it is therefore not totally unexpected that APRc autolytic cleavage sites do not perfectly match the observed specificity preferences of the activated form used in PICS. This observation suggests either a different conformational arrangement of the protease or a dependence on the sequence context and/or conformation of the substrate.

Altogether, the results provided by PICS analysis raise exciting questions about APRc specificity preferences. In fact, the remarkable similarity between the specificity determinants of APRc and other APs, corroborates the findings of Chapter II by providing further experimental validation for the inclusion of APRc in the family of the retropepsin enzymes (A2). Moreover, the detailed information of the specific requirements of APRc here presented might also contribute to the design of new peptide substrates for the development of a highly specific protease-based diagnostic method for the detection of rickettsial infections, and it might also provide a basis for the development of structure-based inhibitors for the treatment of these diseases.

Chapter IV. Functional studies on APRc

Chapter IV. Functional studies on APRc

4.1. Introduction

The emergent and severe character of rickettsioses allied with the lack of protective vaccines, strengthen the importance of identifying new protein factors that may work as potential targets for the development of more efficacious therapies against these diseases[110,111]. Important mediators of bacteria-endothelial cell interaction, like rickettsial adhesins rOmpB and rOmpA, as well as putative mediators of phagosomal escape (e.g., PLD or hemolysin C[198]) or activators of actin-based motility (RickA[200]) have been already identified in *Rickettsia*. However, they are only part of the complex puzzle of host-rickettsial interactions that occurs *in vivo* as many fundamental factors (both in rickettsiae and host cells) still remain undisclosed[139,160,167]. The many limitations imposed by the obligate intracellular lifestyle of *Rickettsia* have hindered the identification of additional rickettsial virulence factors, which would be dramatically aided by a system for genetically manipulating the organism. Therefore, future progress is still required to fully understand the functions of all identified putative virulence factors in pathogenesis, as well as to attribute functions to rickettsial genes that have been annotated without experimental analysis.

In line with this notion the previous two chapters highlighted the identification and biochemical characterization of APRc, a novel retropepsin-like enzyme from *R. conorii.* In this chapter, we start assessing the functional role of APRc to shed light on its role on rickettsial pathogenesis. In fact, we demonstrate that this novel AP is expressed *in vivo* in two pathogenic species of *Rickettsia* and provide experimental evidence for its potential action as a modulator of rickettsial surface cell antigen proteins rOmpB and rOmpA. These results combined with the striking pattern of RC1339/APRc conservation among all rickettsial sequenced genomes strongly suggest that APRc may be an important player in rickettsial pivotal processes. Moreover, because the existence of retropepsin-like enzymes in prokaryotes has always been a matter of debate[16,37], this work also provides the first unequivocally evidence for a retropepsin-like enzyme in gram-negative intracellular bacteria, thereby contributing to change the paradigm on the evolutionary relationships of APs.

4.2. Materials and Methods

Materials

Oligonucleotide primers were purchased from Integrated DNA Technologies, Leuven, Belgium. The synthetic gene encoding the full-length RC1339, the synthetic peptide PeprOmpB (Met-Ala-Gly-Pro-Glu-Ala-Gly-Ala-Ile-Pro-Ala-Ala-Val-Ala-Ala-Gly-Asp-Glu-Ala-Val-Aps-Asn-Val-Ala-Tyr-Gly-Ile-Trp-Ala-Lys), the mouse monoclonal antibody anti-His and the rabbit polyclonal antibody raised towards the sequence Cys-Tyr-Thr-Arg-Thr-Tyr-Leu-Thr-Ala-Asn-Gly-Glu-Asn-Lys-Ala (anti-APRc) were obtained from GenScript (Piscataway, NJ, USA). The plasmid pYC9 and pMC022 encoding *R. conorii* rOmpB and rOmpA, respectively, and anti-rOmpB$_{35\text{-}1334}$ and anti-rOmpA rabbit polyclonal were kindly provided by Dr. Juan Martinez (Louisiana State University, USA). Rabbit polyclonal antibodies raised towards *E. coli* Lap and OmpA (anti-Lep and anti-OmpA, respectively), were generous gifts from Professor Gunnar Von Heine (Stockholm University, Sweden).

DNA constructs

The generation of the constructs encoding the full-length of RC1339/APRc with or without a C-terminal His-tag (pET-APRc$_{1\text{-}231}$-His and pET-APRc$_{1\text{-}231}$, respectively), was previously described in Chapter II, under the section “Materials and Methods, DNA constructs”.

For the generation of the active site mutant pET-APRc(D140A)$_{1\text{-}231}$, the same experimental protocol used for the mutants pGST-APRc(D140A)$_{87\text{-}231}$ and pET-APRc(D140A)$_{99\text{-}231}$-His (Chapter II, section “Materials and Methods, DNA constructs”) was applied.

All positive clones were selected by restriction analysis and confirmed by DNA sequencing.

Full-length APRc expression and *E. coli* cell fractionation

For isolation of total and outer membrane fractions of *E. coli* BL21 Star (DE3) cells expressing full-length rAPRc$_{1\text{-}231}$ the protocol used was essentially as described by Mikado in [279]. BL21 Star (DE3) cells were transformed with pET-APRc$_{1\text{-}231}$ construct and grown at 37 °C until an OD$_{600nm}$ of 0.6-0.7. Expression of rAPRc$_{1\text{-}231}$ was induced with 0.1 mM IPTG for 3 h, after which cells were pelleted by centrifugation at 9000*g* for 20 min at 4 °C and resuspended in PBS buffer. Cells were then mechanically disrupted on a FrenchPress following the manufacturer's instructions (3x, 1500 psi), and cleared by centrifugation at 20000*g* for 20 min. After cell disruption, total membrane fractions were directly pelleted by ultracentrifugation at 144028*g* for 1 h at 4 °C and resuspended in PBS buffer. For enrichment and purification of outer membrane proteins, inner membrane proteins were extracted by incubating the supernatant of lysate clearance with sarkosyl (final concentration of 0.5%) at room temperature for 5 min. Outer membranes were pelleted by ultracentrifugation at 144028*g* for 1 h at 4 °C and resuspended in PBS buffer. Total and outer membrane proteins were resolved by SDS-PAGE and analyzed by immunoblotting using antibodies against APRc, Lep and OmpA, the last two used as internal markers for inner and outer membranes of *E. coli*, respectively.

The protocol for rAPRc$_{1\text{-}231}$-His (pET-APRc$_{1\text{-}231}$-His) and rAPRc(D140A)$_{1\text{-}231}$ (pET-APRc(D140A)$_{1\text{-}231}$) expression and further outer membrane proteins isolation was the same as used for rAPRc$_{1\text{-}231}$. Total and outer membrane proteins were analyzed by immunoblotting using anti-His and anti-APRc antibodies, respectively.

Flow cytometry

E. coli BL21 (DE3) cells were transformed with pET-APRc$_{1\text{-}231}$ construct and grown at 37 °C until an OD$_{600nm}$ of 0.6-0.7. Protein expression was induced with 0.1 mM IPTG for 3 h. Cells were then fixed for 20 min in 4% paraformaldehyde (PFA) and subsequently washed in cold PBS. Fixed cells were incubated with anti-APRc rabbit polyclonal (2 μg/mL) and anti-RNAPα (alpha subunit of *E. coli* RNA polymerase) mouse monoclonal (50 ng/mL) antibodies, and then labeled with both goat anti-rabbit IgG Alexa Fluor 488 (Life Technologies) and goat anti-mouse IgG R-PE-Cy5.5 conjugated secondary antibodies (SouthernBiotech) at the specified concentration (4 μg/mL). Bacteria were analyzed by flow cytometry using a BD FACSCalibur (BD Biosciences) instrument and FlowJo software. For analysis of non-permeable *E. coli* cells, positive anti-RNAPα staining cells were gated out, and intact cells analyzed for surface expression of rAPRc$_{1\text{-}231}$ with anti-APRc antibody.

RT-PCR analysis

For cDNA synthesis, total RNA was isolated from an aliquot of frozen *R. conorii* Malish 7 and *R. rickettsii* "Sheila Smith", *R. rickettsii* Iowa, *R. parkeri* Portsmouth, *R. montanensis* OSU 85-930, *R. amblyomii* GAT-30V-infected Vero cells and *R. felis* URRWXCal2-infected ticks using the SurePrep TrueTotal RNA Purification Kit (Fisher Scientific), according to the manufacturer's instructions. After extraction, RNA samples were treated with DNase I, RNase-free set (Thermo Scientific) for 30 min at 37 °C. The reaction was inactivated by adding 50 mM EDTA and heating the mixture at 65 °C for 10 min. Next, approximately 1 μg of total RNA was used as the template for reverse transcription using the iScript™ cDNA Synthesis Kit (BioRad), according to the manufacturer's instructions. For all extracted samples, negative RT-PCR controls were processed in the absence of reverse transcriptase. APRc gene expression was assessed by PCR reaction with the specific primers RC1339_RT-Fwd (5'-AAAGCCGCCCCTATAACCTT-3') and RC1339_RT-Rev (5'-TCCTGAAACCTTTGAAACGCTC-3') which were designed for the amplification of a segment with 136 bp. The PCRs were performed in a 50 µl volume, with 1 µl of cDNA as the DNA template, 0.1 µM of each primer, 1x PCR buffer (100 mM Tris-HCl (pH 9.0), 15 mM $MgCl_2$, 500 mM KCl), 200 µM of dNTP mix, and 1 U of *Taq* DNA polymerase (GE Healthcare). The PCR mixtures were incubated at 95 °C for 3 min, followed by 35 cycles of 95 °C (30 sec), 55 °C (30 sec), and 72 °C (30 sec). The gene *hrtA* (17 kDa surface antigen) was used as the positive control using the primers Rc_htrA_RT-Fwd (5'-GGACAGCTTGTTGGAGTAGG-3') and Rc_htrA_RT-Rev (5'-TCCGGATTACGCCATTCTAC-3'). An aliquot of 20 µl of each PCR product was electrophoresed on a 1.7% agarose gel and stained with ethidium bromide. The size of the PCR product was determined by comparison with GeneRuler^{TM} 1 kb Plus DNA Ladder (Thermo Scientific).

R. conorii and *R. rickettsii* cell fractionation

Cell fractionation studies with *Rickettsia* spp. were performed as previously described in [280]. Briefly, approximately $5x10^6$ plaque forming units (pfu) of purified *R. conorii* Malish 7 or *R. rickettsii* "Sheila Smith" was fixed in 4% PFA in PBS, washed in PBS and then removed from BSL3 containment after verification that viable rickettsiae were no longer present according to standard operating procedures. For whole-cell lysates, cells were resuspended in SDS-PAGE loading buffer and boiled. Total outer membrane proteins were extracted essentially as described by Nikaido in [279]. The sample was resuspended in 1.5 mL of 20 mM Tris-HCl pH 8.0 containing 1X protease inhibitor cocktail and then subjected to three rounds of French press treatment for cell lysis. The resulting lysate was centrifuged at 10000*g* for 3 min to remove unbroken cells and then incubated in 0.5% sarkosyl at room temperature for 5 min. The sarkosyl-treated lysate was centrifuged at >16000*g* for 30 min at 4 °C. The sarkosyl soluble protein fraction was removed and the remaining insoluble pellet representing the outer-membrane protein fraction was washed in 20 mM Tris pH 8.0 and then boiled in 0.5 mL of 20 mM Tris-HCl pH 8.0 containing 0.5 mL of 2X SDS sample buffer. Protein samples were aliquoted and frozen at -20 ˚C until use.

rOmpB expression and trans-activation assays

The constructs encoding *R. conorii* rOmpB (pYC9) and rOmpA (pMC022) and their expression in *E. coli* was performed as previously described in [186] and [183], respectively. *E. coli* BL21 (DE3) cells transformed with pYC9 or pMC022 plasmids were grown in LB medium at 37 ˚C supplemented with ampicillin. Bacteria were diluted 1:20 from overnight cultures, grown to an OD_{600nm} of 0.6-0.7, and induced with 0.1 M IPTG for 3 h at 30 ˚C. After expression, the cells were harvested by centrifugation at 36000*g* for 20 min, at 4 ˚C and the pellet resuspended in 20 mM Tris-HCl pH 8 before freezing at -20 ˚C.

To assess for *in vitro* proteolytic cleavage of these outer membrane proteins by APRc, the total membrane fraction of *E. coli* cells overexpressing either rOmpB or rOmpA were isolated as described under the section “Full-length APRc expression and *E. coli* cell fractionation”, and then incubated for 16 h at 37 ˚C in 50 mM sodium acetate pH 4.0, 100 mM NaCl with 25 µg of purified active APRc ($APRc_{105-231}$-His). A parallel incubation was performed under similar conditions with the active site mutant $rAPRc(D140A)_{99-231}$-His as a negative control. The reaction products were separated by SDS-PAGE and analyzed by Western blot with anti-APRc, anti-$rOmpB_{35-1334}$ and anti-rOmpA rabbit polyclonal antibodies.

PeprOmpB cleavage assays by APRc

To further validate the ability of APRc to process rOmpB between the passenger domain and the beta-barrel translocation domain, a synthetic peptide with the sequence Met-Ala-Gly-Pro-Glu-Ala-Gly-Ala-Ile-Pro-Ala-Ala-Val-Ala-Ala-Gly-Asp-Glu-Ala-Val-Aps-Asn-Val-Ala-Tyr-Gly-Ile-Trp-Ala-Lys (PeprOmpB) was synthetically synthesized. This peptide comprises the predicted cleavage region in rOmpB from *R. conorii*, as determined for *R. typhi* and *R. prowazekii*[187] rOmpB proteins. For the incubation assay, 2.5 µg of active APRc ($APRc_{105-231}$-His) were added to 100 µg of PeprOmpB in 0.1 M Sodium Acetate pH 6.0 for 16 h at 37 ˚C. To evaluate the inhibitory effect of indinavir and nelfinavir on the ability of APRc to cleave PeprOmpB, similar incubation assays were performed under the presence of each inhibitor at 1 mM final concentration. Parallel incubations performed under the same conditions but in the absence of APRc were used as control. The reaction mixtures were then centrifuged at 20000*g* during 6 min and the digestion fragments separated by RP-HPLC on a C18 column (KROMASIL 100 C18 250, 4.6 mm), using a Prominence system (Shimadzu Corporation, Tokyo, Japan). Elution was carried out with a linear gradient (0–80%) of acetonitrile in 0.1% v/v trifluoroacetic acid for 30 min at a flow rate of 1 mL/min. Absorbance was monitored at 220 nm. Each corresponding control spectrum was used as baseline for spectral subtraction. Selected eluted peaks were collected and analyzed by LC-MS/MS (Center for Neuroscience and Cell Biology, Proteomics Unit).

SDS-PAGE and Western blotting

SDS-PAGE analysis was performed in a Bio-Rad Mini Protean III electrophoresis apparatus using 4-20% or 12.5% polyacrylamide gels. Samples were treated with loading buffer (0.35 M Tris-HCl, 0.28% SDS buffer pH 6.8, 30% glycerol, 10% SDS, 0.6 M DTT and 0.012% Bromophenol Blue) and boiled for 5 min before loading. Gels were stained with Coomassie Brilliant Blue R-250 (Sigma). For Western blot analysis, protein samples were resolved by SDS-PAGE and electrotransferred onto PVDF or nitrocellulose membranes by standard wet (using the buffer 25 mM Tris, 192 mM glycine and 20% methanol) or semi-dry (in buffer 25 mM Tris, 192 mM glycine, 20% methanol and 0.025% SDS) transfer apparatus. Membranes were then blocked for one hour in standard TBS containing 1% (v/v) Tween-20 supplemented with 5% (w/v) skim milk or 2% (w/v) BSA and then incubated with anti-APRc (GenScript, 2 µl/mL), anti-rOmpB$_{35-1334}$ (*R. conorii*), anti-rOmpA (*R. conorii*) rabbit polyclonal antibodies, anti-OmpA (*E. coli*) serum and anti-Lep (*E. coli*) serum. Membranes were washed in TBS, containing 0.1% (v/v) Tween-20 and incubated with secondary anti-mouse or anti-rabbit alkaline phosphatase-conjugated (GE Healthcare) and IRDye-conjugated (LI-COR Biotechnology) antibodies and revealed using ECF chemiluminescence detection kit (GE Healthcare) in a Molecular Imager FX (Bio-Rad) or by infrared detection using an Odyssey infrared dual-laser scanning unit (LI-COR Biotechnology), respectively.

To confirm the specific band reactivity of anti-APRc antibody, a peptide competition assay was performed. The primary antibody was pre-incubated with 100-fold (mass) excess of immunizing peptide (CYTRTYLTANGENKA) for 20 min at room temperature prior immunoblotting analysis and parallel experiments were performed with pre-incubated and non-incubated antibody.

4.3. Results

Full-length APRc expression and subcellular localization studies in *E. coli* cells

As previously stated, a major difference between full-length APRc (rAPRc$_{1-231}$) and retroviral proteases is the predicted membrane-embedded N-terminal domain of APRc and C-terminal domain with an extracytoplasmic orientation. In order to provide experimental validation of these theoretical observations we used *E. coli* as our working model. An untagged construct in pET28a comprising RC1339/APRc full-length coding sequence was generated and protein expression carried out as described under experimental procedures. After optimizing expression conditions to achieve the best yields of rAPRc$_{1-231}$, sub-cellular fractionation studies with sarkosyl followed by Western blot analysis with the specific APRc antibody were undertaken to assess the insertion of this protein into the membrane and determine its location (inner vs. outer membrane). Sarkosyl, also known as sodium lauroyl sarcosinate, is an anionic amphiphilic surfactant due to the hydrophobic 14-carbon chain (lauroyl) and the hydrophilic carboxylate. Sarkosyl treatment is commonly used in the purification of outer membrane proteins of gram-negative bacteria given its ability to selectively solubilize inner membrane proteins and to produce the purest and most reproducible preparations of outer membrane proteins[281]. Because recombinant membrane proteins have a high tendency to aggregate, the use sarkosyl has also the advantage of solubilizing proteins in the form of inclusion bodies, greatly reducing the potential contamination of outer membrane fractions[282].

From this immunoblotting analysis resulted the identification of a band with approximately 21 kDa in the total membrane fraction, shown to accumulate in the outer membrane (Figure 26A). The nature of this signal was further confirmed by peptide competition assays (Figure 26A, right panel). The purity of the outer membrane fraction was confirmed by Western blotting against *E. coli* Lep and OmpA proteins, as inner and outer membrane markers[283], respectively, and compared to the total membrane fraction (Figure 26B). As expected, *E. coli* OmpA was detected in the outer membrane fraction and the absence of cross-contamination with inner membrane proteins was confirmed through loss of signal for Lep, when compared with total membrane fraction. Interestingly, APRc displayed a molecular weight lower than expected (~21 kDa instead of the predicted 26.4 kDa), and parallel experiments with a C-terminal His-tagged construct (rAPRc$_{1-231}$-His) confirmed the presence of the tag in the membrane fractions (Figure 26C, left panel), clearly suggesting that the protease may be processed at the N terminus during translocation to the outer membrane. Moreover, when the active site mutant of full-length APRc (rAPRc(D140A)$_{1\text{-}231}$) was expressed under the same conditions, the same apparent molecular weight was observed, thereby ruling out a potential auto-proteolytic event (Figure 26C, right panel).

In an attempt to expand our knowledge about the membrane topology of APRc, further studies were performed in order to determine the overall in/out orientation of this protein relative to the outer membrane of *E. coli*. To investigate this, PFA-fixed *E. coli* cells expressing untagged full-length APRc were subjected to flow cytometry with both anti-APRc and anti-RNAPα antibodies. The staining of *E. coli* cells with the RNAPα mAb was primarily used to restrict the analysis to the non-permeable cells. As shown in Figure 26D, after gating out all the cells that stained positive for RNAPα (permeable cells), bacterial surface staining with anti-APRc was observed, confirming the integration of RC1339/APRc into the outer membrane of *E. coli* and the orientation of the soluble catalytic domain to the extracellular milieu.

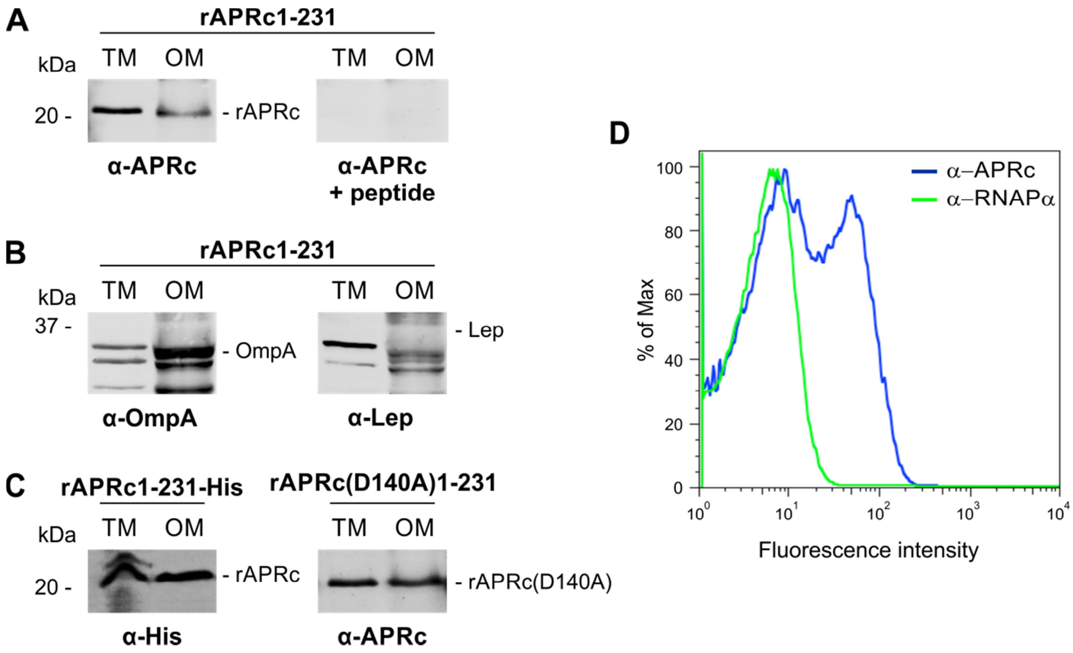


Figure 26. *Recombinant full-length APRc localization in E. coli and cell surface orientation of the soluble catalytic domain.* (**A**) Full-length APRc was expressed in *E. coli* and total (TM) as well as outer membrane (OM) fractions were isolated and analyzed by Western blot with anti-APRc antibody (left panel). As a control for non-specific staining, peptide competition assays were performed by blocking the anti-APRc antibody with the immunizing peptide (right panel). One specific band with approximately 21 kDa detected in the outer membrane fraction reveals that recombinant full-length APRc accumulates in this *E. coli* membrane. (**B**) The purity of outer membrane fractions was confirmed by using OmpA and Lep proteins as internal markers for the outer and inner membranes of *E. coli*, respectively. Both proteins were present in TM faction, while only OmpA is detected in outer membrane faction. (**C**) The positive signal observed for TM and outer membrane fractions of *E. coli* cells expressing rAPRc$_{1\text{-}231}$-His confirmed that the C-terminus is kept intact, thereby suggesting that APRc might be proteolytic processed at the N terminus (left panel). The same molecular weight of approximately 21 kDa observed for APRc(D140A)$_{1\text{-}231}$, strengthens the assumption of an N-terminal cleavage by the translocation machinery (right panel). (**D**) Flow cytometric analysis was carried out for recombinant APRc recognition at the surface of *E. coli* cells. PFA-fixed *E. coli* cells were incubated with anti-APRc and anti-RNAPα, followed by secondary detection using goat anti-rabbit IgG Alexa Fluor 488- and goat anti-mouse IgG R-PE-Cy5.5 conjugated secondary antibodies, respectively. After gating out the subpopulation of cells staining positive for RNAPα (permeable cells), fluorescence was detected on *E. coli* cells incubated with anti-APRc, thereby confirming the expression of recombinant APRc at the outer membrane and its exposure to extracellular milieu.

APRc transcriptional analysis and localization studies in *Rickettsia*

To determine whether *RC1339* and its homologues from other five rickettsial species are expressed in the context of intact *Rickettsia* cells, we isolated the total RNA from *R. conorii* Malish 7 and *R. rickettsii* "Sheila Smith", *R. rickettsii* Iowa, *R. parkeri* Portsmouth, *R. montanensis* OSU 85-930, *R. amblyomii* GAT-30V grown in Vero cells and *R. felis* URRWXCal2 – from infected ticks cells, and performed reverse transcriptase PCR (RT-PCR). As shown in Figure 27, *R. conorii*, *R. rickettsii*, *R. montanensis* and *R. felis* produce the transcript for the retropepsin-like enzyme. Moreover, despite the apparent lack of gene amplification for *R. parkeri* and *R. amblyomii*, one cannot rule out the transcription of this gene in these two species. In fact, since an equal amount of RNA was used in cDNA synthesis, the lower transcription levels of *hrtA* gene in comparison with those of the other four analyzed species, suggests a substantial contamination with RNA from Vero cells on those samples. Due to the low sensitivity of conventional RT-PCR analysis, the reduced amount of bacterial RNA used in the PCR reaction may have been manifestly insufficient for the detection of APRc transcripts.

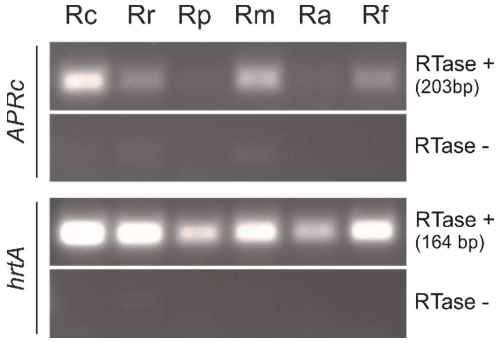


Figure 27. *RT-PCR analysis of APRc expression on rickettsial spp.* Gene expression of APRc and correspondent homologues was assessed by conventional RT-PCR for six rickettsial species (Rc: *R. conorii* Malish 7, Rr: *R. rickettsii* Iowa, Rp: *R. parkeri* Portsmouth, Rm: *R. montanensis* OSU 85-930, Ra: *R. amblyomii* GAT-30V, Rf: *R. felis* URRWXCal2). The housekeeping gene hrtA (17 kDa surface antigen) was used as a control. The amplification product of APRc gene was evident for Rc, Rr, Rm and Rf, although lower levels of expression are observed for Rr and Rf in relation to Rc and Rm (when compared to the expression levels of the hrtA gene). The apparent lack of APRc gene amplification for Rp and Ra is likely due to the low amount of mRNA, as indicated by the low amplification levels of hrtA gene. The negative control for the cDNA synthesis lacking reverse transcriptase is identified by (RTase -). Rickettsial species are identified on the top and the gene names are shown on the left side of the agarose gel.

Then, protein lysates from *R. conorii* and *R. rickettsii* were separated by SDS-PAGE and immunoblotting analyses were carried out with the specific APRc antibody. As depicted in Figure 28, a major reactive species with an apparent molecular mass of 21 kDa was detected in *R. rickettsii* whole cell lysate and in the insoluble fraction of the *R. conorii* extract. These results clearly confirmed that *RC1339* gene and its *R. rickettsii* homologue are indeed translated in both rickettsial species. Interestingly and as previously observed in *E. coli*, a molecular weight of around 21 kDa was also detected for APRc in rickettsial extracts. Although we cannot exclude abnormal migration of the protease in the gel, the observed lower molecular weight may also be correlated with APRc processing at the N terminus, as anticipated by our results in *E. coli*.

To provide additional insights on the localization of APRc in these rickettsial species, fractionation studies using sarkosyl treatment were also carried out on purified bacteria.

Whole cell lysates as well as isolated inner and outer membrane fractions were separated by SDS-PAGE and analyzed by Western blot. For both species tested, our results were consistent with localization of the protease at the outer membrane, as confirmed by the immunodetection of rickettsial rOmpB which was used as an internal marker for the outer membrane in these assays (Figure 28B, bottom panel). This further corroborates the localization of APRc at the outer membrane of rickettsial species. Taken together, we have shown that this novel retropepsin-like enzyme is expressed *in vitro* in two pathogenic species of *Rickettsia* and, furthermore, provide evidence for its localization at the outer membrane of these bacteria.

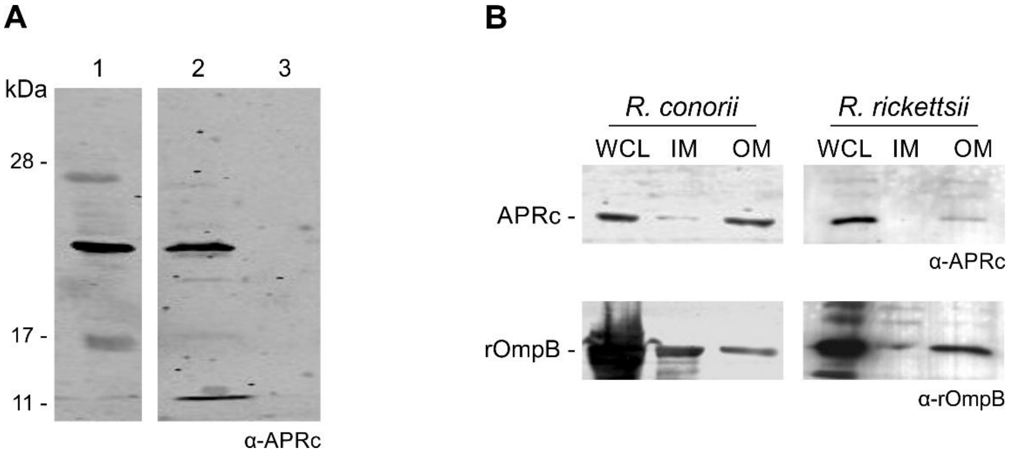


Figure 28. *APRc in vivo expression in R. conorii and R. rickettsii and outer membrane localization.* (**A**) A whole cell lysate from *R. rickettsii* (1) and insoluble (2) and soluble (3) fractions from *R. conorii* extracts were isolated and then subjected to Western blot analysis with anti-APRc antibody. A specific band with approximately 21 kDa was detected. (**B**) Whole cell lysates (WCL), inner (IM) and outer membrane (OM) fractions from sarkosyl treatment of *R. rickettsii* and *R. conorii* extracts were isolated and then subjected to Western Blot analysis with anti-APRc and anti-rOmpB antibody. APRc shares the same localization of rOmpB, an internal marker for outer membrane of *Rickettsia* spp. Molecular weight markers in kilodaltons (kDa) are shown on the left.

rOmpB trans-activation assays with active APRc

The evidence that a proportion of APRc is associated with the outer membrane led us to hypothesize that rickettsial surface proteins might be potential substrates for this newly characterized enzyme. As has been shown for other autotransporter proteins, rickettsial rOmpB, rOmpA, Sca1, and Sca2 proteins are involved in mediating important interactions with mammalian cells and undergo processing events at the outer membrane[184–188,217,284]. As an example, *R. conorii* rOmpB is expressed as a preprotein of 168 kDa and is subsequently cleaved to release the passenger domain (120 kDa) from the β-barrel translocation domain (32 kDa)[187].

Interestingly, *R. conorii* and *R. japonica* rOmpB do not undergo proteolytic cleavage when expressed at the outer membrane of *E. coli,* suggesting that the processing event is not autocatalytic[186]. However, the identity of the enzyme responsible for Sca protein maturation still remains elusive. Therefore, and based on the observed APRc outer membrane localization, we sought to determine whether APRc might participate in the processing of rOmpB (Figure 29A). In order to do this, we performed transactivation assays using *E. coli* outer membrane fractions enriched in recombinant rOmpB (C-terminally His-tagged) and purified active APRc (soluble catalytic domain). Reaction products were then separated by SDS-PAGE and analyzed by Western blot. As shown in Figure 29B, the detection of an anti-His immune reactive product with ~35 kDa in the presence of APRc was correlated with the disappearance of rOmpB proprotein, suggesting that this enzyme may be indeed capable of promoting cleavage of recombinant rOmpB. Moreover, the generated reactive protein product has approximately the same molecular weight as that expected for rOmpB β-barrel (32 kDa), further suggesting that this proteolytic cleavage may likely be occurring somewhere between the passenger and the β-barrel domain, in agreement with what has been described for native rOmpB[187]. To further validate these results, parallel assays were performed in the presence of APRc active site mutant and the integrity of rOmpB proprotein evaluated by immunoblotting with a specific antibody to this outer membrane protein (Figure 29C). As expected, the disappearance of rOmpB proprotein was observed in the presence of active APRc but not when the cell extract was incubated with the active site mutant protein (APRc(D140A)$_{99\text{-}231}$-His).

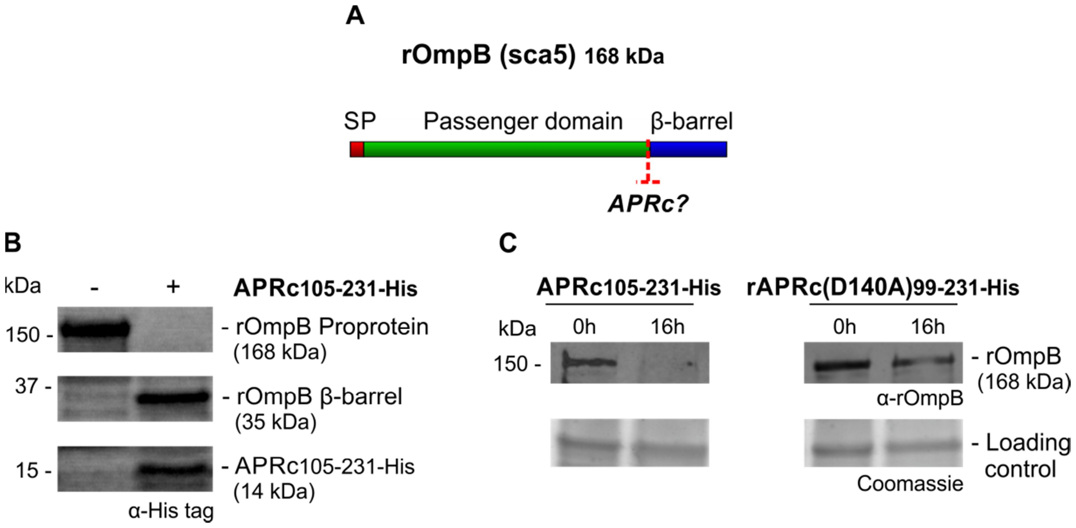


Figure 29. *APRc can process rickettsial OmpB in vitro.* (**A**) rOmpB is proteolytically processed between the passenger and β-barrel domains through a yet unknown mechanism (?) and APRc was tested as the candidate enzyme to perform rOmpB proprotein processing *in vitro*. (**B**) Total membrane fractions of *E. coli* enriched in rOmpB were incubated with activated APRc soluble domain and the reaction products analyzed by Western blot with an anti-His antibody. The integrity of rOmpB proprotein was confirmed in the absence of APRc whereas in the presence of the protease a product with approximately 35 kDa was observed, correlated with the disappearance of the full-length unprocessed form. (**C**) The integrity of recombinant rOmpB was further evaluated upon incubation with both activated APRc and the active site mutant form (D140A) for 16 h. The reaction products were then subjected to immunoblot analysis with anti-rOmpB, confirming the disappearance of rOmpB in the presence of the active form of the enzyme. Molecular weight markers in kilodaltons (kDa) are shown on the left. Protein loading controls: Coomassie blue staining.

Taking into account the ability of APRc to cleave rOmpB, we decided to extend this analysis to another conserved rickettsial antigen, rOmpA. Interestingly, a similar phenomenon was observed, demonstrating that a protein other than rOmpB can be processed by APRc *in vitro* (Figure 30).

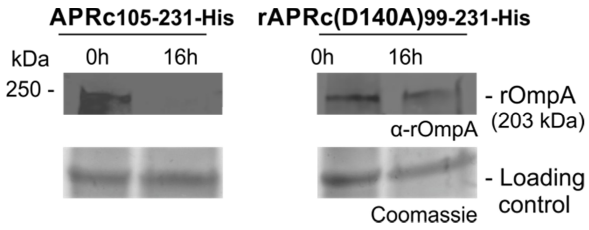


Figure 30. *APRc can process rOmpA in vitro*. Total membrane fractions of *E. coli* enriched in rOmpA were incubated with both activated APRc and the active site mutant form (D140A) for 16 h. The reaction products were then subjected to immunoblot analysis with anti-rOmpA, confirming the disappearance of rOmpA in the presence of the active form of the enzyme. Molecular weight markers in kilodaltons (kDa) are shown on the left. Protein loading controls: Coomassie blue staining.

Altogether, these results suggest that APRc is sufficient to mediate rOmpB and rOmpA maturation *in vitro*, thereby raising an exciting hypothesis regarding possible functional significance of APRc as being able to process these and possibly other autotransporter proteins in the context of intact *R. conorii* cells. Therefore, whether other Sca proteins can also serve as substrates might be the basis for subsequent studies.

Cleavage of a synthetic rOmpB peptide by APRc

Given the promising results obtained with the trans-activation assays of rOmpB, and in order to further confirm the ability of APRc to cleave this outer membrane protein and determine the site of cleavage, a synthetic peptide corresponding to the sequence between the passenger and the β-barrel domains of *R. conorii* rOmpB (Met-Ala-Gly-Pro-Glu-Ala-Gly-Ala-Ile-Pro-Ala-Ala-Val-Ala-Ala-Gly-Asp-Glu-Ala-Val-Asp-Asn-Val-Ala-Tyr-Gly-Ile-Trp-Ala-Lys) was synthetized. The peptide was incubated with active APRc for 1 h, and the reaction products were then separated by RP-HPLC. As shown in Figure 31 it was possible to confirm activity towards this peptide by the appearance of three major peaks in the chromatogram (#1, #2 and #3). Moreover, and as shown in Figure 31A, in the presence of indinavir (left panel) and nelfinavir (right panel) the activity towards this substrate was inhibited, further confirming the specificity of APRc and corroborating our previous observations in trans-activation assays. The subsequent analysis by LC-MS/MS of these three eluted fractions, led to the combined identification of different cleavage sites within the region Ala*Val*Ala*Ala*Gly*Asp*Glu-Ala*Val*Asp*Asn (* corresponded to identified cleavage sites). Although the identification of several cleavage sites may result from the fact that enzyme and substrate were incubated for one hour, this result is nevertheless very exciting and consistent with the region spanning the cleavage site, which has been experimentally determined through the N-terminal sequencing of the *R. typhi* and *R. prowazekii* rOmpB β-peptide[187], corresponding to the sequence Ala-Ala-Val-Ala-Ala*Gly-Asp-Glu-Ala-Val (*, cleavage site identified by Edman degradation). Improvements on future assays, including shorter incubation times and quantitative analysis, will be required to identify the preferential cleavage site within PeprOmpB as well as the most abundant reaction product for each identified peak.

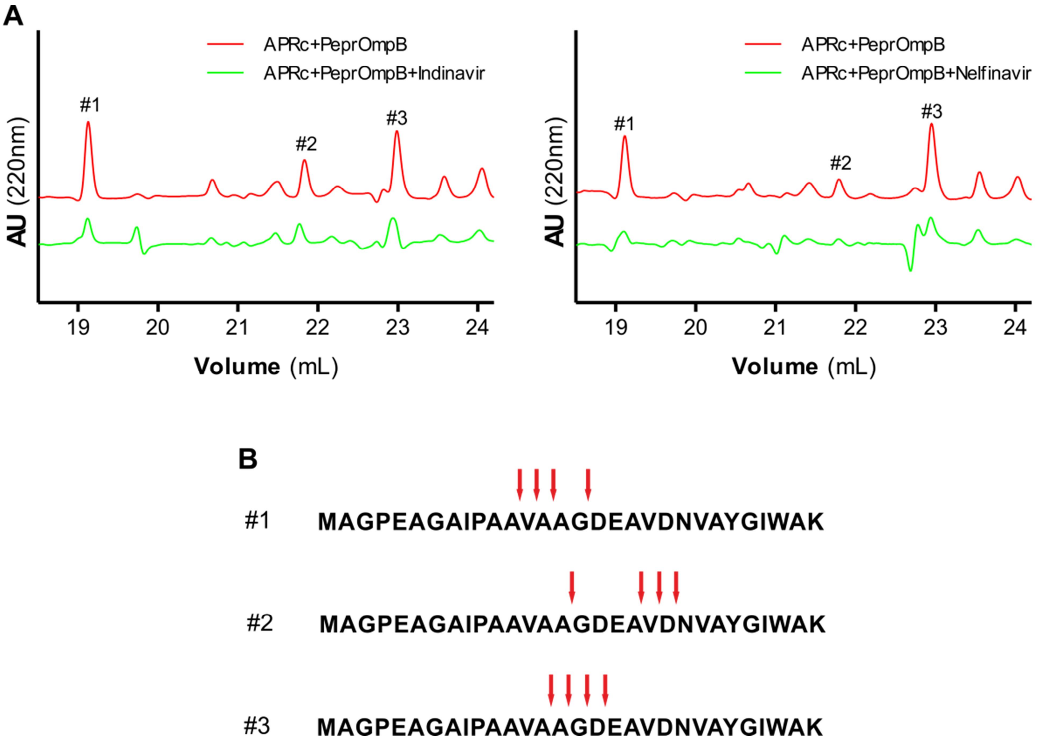


Figure 31. *PeprOmpB cleavage by APRc and its inhibition by indinavir and nelfinavir.* (**A**) Reaction products from the incubation of PeprOmpB (a synthetic peptide corresponding to the sequence between the passenger and the β-barrel domains of *R. conorii* rOmpB) with APRc under the presence or absence of indinavir (left panel) and nelfinavir (right panel) were separated by RP-HPLC and the major peaks (#1, #2 and #3) collected for further analysis by LC-MS/MS. (**B**) Identified cleavage sites for each peak within the region Ala-Val-Ala-Ala-Gly-Asp-Glu-Ala-Val-Asp-Asn are indicated by red arrows.

4.4. Discussion

The increasing number of sequenced pathogenic bacterial genomes is producing an ever expanding gap between sequence data and the efficient identification of genes required for pathogens to cause disease. Although comparative genomics and other genomic tools have been instrumental in the identification of numerous pathogen genes, understanding how these genes contribute for disease requires the validation of their expression and the determination of their molecular function[285].

In this work, we provide the first evidence for the expression of APRc both at the transcriptional and translational levels in *Rickettsia* spp.. Importantly, we also demonstrate that *APRc* expression is not confined to recognized pathogenic species such as *R. conorii* and *R. rickettsii*, but also expressed in *R. montanensis*, whose role in human health has yet to be determined. In line with these evidences, Bechah and colleagues[160] reported the transcription of *RC1339* gene homologue (*RP867*) in *R. prowazekii*. Notably, the comparison of the transcriptional profiles obtained for *R. prowazekii* Rp22 (virulent) and Erus (avirulent) strains has also revealed a differential regulation of *RP867* gene expression with a fold change of 1.77 between the virulent and the avirulent strains[160]. Because bacterial virulence is a multifactorial process, careful speculations should be made concerning a possible direct role of APRc for rickettsial virulence. Therefore, it will be critical to evaluate whether similar differential transcript abundance of *RC1339/APRc* gene can be observed among other rickettsial species with different degrees of virulence, complemented by studies on the dynamics of transcript levels during the different stages of infection. Besides the assessment of gene expression of APRc by quantitative RT-PCR (qRT-PCR), a comparative immunoblot analysis of protein expression, localization and auto-processing activity in different rickettsial species at different time points post infection, will allow us to determine a possible relationship between expression, patterns of secretion and production of APRc processed forms, with the degree of pathogenicity/virulence of different rickettsial strains.

Sub-cellular localization studies have also revealed an outer membrane accumulation for APRc in *Rickettsia* spp. which was also confirmed by expression of the full-length protease in *E. coli*. Gram-negative bacteria contain two lipid bilayers which differ markedly with respect to composition and function, the inner and outer membrane. The asymmetric outer membrane is composed of lipopolysaccharides in the outer leaflet and phospholipid in the inner leaflet while the symmetric inner membrane is composed of a phospholipid bilayer[286]. Both membranes contain proteins which are crucial players in the cell, and take center stage in processes ranging from basic small-molecule transport to sophisticated signaling pathways, and many of them are important drug targets[287]. In particular, the outer membrane proteins are known to play essential roles in energetics, metabolism, signal transduction and transport[288]. Despite the recognized importance of membrane proteins, the high-resolution three-dimensional structures of membrane proteins are still very hard to obtain, representing less than 1% of the structures in the Protein Data Bank[289]. Bacterial membrane proteins can be either in the form of integral membrane proteins or as lipoproteins that are anchored to the membrane by means of N-terminally attached lipids. Among all integral membrane proteins, two basic architectures are recognized: the α-helical type which is the most abundant and occurs mostly in inner membranes, and the β-barrels which are known from outer membranes of bacteria[290,291]. The α-helices of inner membrane proteins are composed by continuous stretch of 20-30 nonpolar residues that cross the membrane, with a predominance of aliphatic side chains at the center and aromatic residues (Trp and Tyr) at both ends. Although having common surface characteristics, the secondary structure and fold of the outer membrane proteins are completely different from those of inner membrane proteins. As the name imply, β-barrel proteins are comprised of β-barrel motifs composed of alternate polar and non-polar amino acids which form mono, di and trimeric structures with 8–22 β-barrels. The non-polar amino acids point into the lipid and protein interface, while the polar amino acids point into the interior of the barrel[292].

From a bioinformatics search through a set of protein sequences derived from genomic DNA sequences, it became evident that, while the helix bundle represents about 20% to 25% of all open reading frame, the β-barrel form represents a few percent of all open reading frame[293]. At the present, transmembrane β-barrel proteins have been found exclusively in the outer membrane of gram-negative bacteria, and these membranes appear to lack α-helical proteins. In view of this, the outer membrane localization of APRc is somehow unexpected. However, at least another transmembrane protein with α-helical architecture have been also reported to be embedded in the outer membrane of gram-negative bacteria[294]. Wza is an integral outer membrane protein that is essential for capsular polysaccharides export. This protein assembles to an octamer with a novel α-helical barrel transmembrane region which forms an elongated cylindrical structure with a molecular weight of 340 kDa. The helices within the barrel are amphipathic to permit interaction with the outer membrane on one hand and the export of polysaccharides on the other hand.

Our results provide additional evidence that the bacterial surface is not restricted to proteins with β-barrel structures[294,295], further suggesting that the repertoire of proteins with α-helices localized to the outer membrane of gram-negative bacteria may be higher than anticipated. Although numerous systems for protein export have already been uncovered in gram-negative bacteria (eight types are known to mediate export across or insertion into the inner membrane, while eight specifically mediate export across or insertion into the outer membrane), none of them has been identified as being capable to export and insert α-helical proteins at the outer membrane[290]. Gram-negative bacterial outer membrane proteins are in general translocated into the periplasm via the Sec pathway, consisting of the SecYEG membrane-spanning translocase complex. Nevertheless, the regulatory network behind outer membrane biogenesis appears to be rather complex and we are far from understanding the molecular mechanisms behind the integration of proteins into the outer membrane. Therefore, one can speculate that a novel outer membrane protein secretion system, yet to be identified, might be implicated on the insertion of α-helical type of outer membrane proteins, such as Wza and APRc, at the surface of gram-negative bacteria. Comparisons among the different bacterial protein secretion systems suggest that each evolved independently despite there are a number of specific biogenic, mechanistic, and evolutionary similarities among them[290].

The conservation of many proteins from translocase machineries between bacteria and mitochondria provides important insights into the evolution of the outer membranes and the development of their protein biogenesis system[296]. Mitochondrial protein translocases, which possibly derived from protein translocation systems in α-proteobacteria, appear to have evolved to their current levels of complexity during or after the degeneration of endosymbiotic bacteria into mitochondria. Interestingly, the mitochondrial outer membrane contains both β-barrel and α-helical proteins. While the translocase of the outer membrane (TOM) complex forms the entry for most nuclear-encoded mitochondrial proteins, the sorting and assembly machinery (SAM) complex is not only essential to the biogenesis of some β-barrel proteins, but is also required for the assembly of a subset of single or multi-spanning α-helical proteins, including some α-helical transmembrane domain proteins from the TOM complex. The SAM core complex was shown to form two major complexes with Mim1 (mitochondrial import protein 1) and Mdm10 (mitochondrial distribution and morphology protein) proteins, with different functions in the biogenesis of α-helical proteins[297,298]. Despite distantly related, Sam50 from SAM complex is conserved from gram-negative bacteria to mitochondria, and homologues (BamA family) are present in virtually all gram-negative bacteria (e.g., YaeT in *Escherichia coli*[299], Omp85 in *Neisserial* spp.[300] and the protective surface antigen D15 in *Haemophilus influenzae*[301]). Proteins from BamA family are core components for the biogenesis of the outer membranes of bacteria as they are required for the effective insertion of lipids and integral proteins into these bacterial membranes[302]. A BLAST search of the amino acid sequence of Sam50 returns an outer membrane protein annotated in rickettsial spp. genome named Omp1. Importantly, although Sam50 and Omp1 only share 25-27% identity at the amino acid level, the latter is highly conserved in *Rickettsia* spp.. Therefore, considering the well-known phylogenetic relationship of *Rickettsiales* and mitochondria, this raises the exciting hypothesis that Omp1 may be part of a new complex responsible for the insertion machinery of α-helical proteins like APRc into the outer membrane.

It is noteworthy that, although no cleavable signal sequence was predicted for APRc, both wild-type APRc$_{1-231}$ and corresponding inactive site mutant (APRc(D140A)$_{1-231}$) display the same molecular weight of approximately 21 kDa (Figure 26). This result either suggests that this protein may have different gel mobility or that an N-terminal sequence is cleaved off. Similar to the function exerted by known bacterial signal peptidases[303], one cannot exclude at this point that the translocation machinery responsible for the transport and insertion of APRc into the outer membrane may be also implicated in this (apparent) N-terminal processing. Nevertheless, these preliminary results definitely require further studies in order to clarify whether the detected 21 kDa band is an intermediate processed form or the result of different gel mobility.

The mechanism by which multi-spanning membrane proteins with α-helix topology are inserted into their target membrane remains to be fully elucidated. In contrast, several factors are known to influence the final transmembrane orientation of these proteins, namely the cooperative action of topogenic sequences, the interactions during folding within the protein and the interactions between the protein and the insertion/translocon machinery and the lipid environment[304–306]. Among these, positively charged residues (Arg and Lys) within the loops flanking the hydrophobic stretches are considered the major topological determinants of membrane proteins, and in particular, it has been statistically derived and experimentally confirmed that these type of residues are four times more abundant in their cytoplasmic side as compared to extra-cytoplasmic domains[304,307]. Although this particular rule, known as the “positive inside rule”[307], has been applied solely to α-helical proteins spanning the inner membrane, our preliminary biochemical results obtained by flow-cytometry analysis also suggests that APRc topology may follow similar principles of insertion at the outer membrane. The observation of a C terminus (catalytic domain) facing the exterior of the bacterial cell with respect to the plane of the outer membrane, implies a periplasmic orientation for the N-terminal of APRc. In good agreement and according to the predicted secondary structure of APRc, the basic residues Arg and Lys are disproportionately favored in the periplasmic cap region, despite some of these residues are also expected in the first loop between TMH1 and

TMH2, which is exposed at the surface of bacteria. Of note, as previously reported for inner membrane proteins, individual transmembrane helices do not always insert into the membrane in a strict N- to C-terminal order but can reorient during the insertion process[304,306,308]. These and other similar observations raised important questions regarding flexibility in the way multi-spanning membrane proteins are handled by the insertion machinery. As observed for EmrE[306], proteins might experience a post-insertion conversion between different topologies, but whether they occur only under extreme conditions or exist naturally in wild-type cells, remains to be elucidated. Similarly, further studies are definitely required to assess if APRc has a fixed topology or a dual-topology which could be seen as a regulatory mechanism to control APRc activity.

Furthermore, if considering the membrane embedded nature of APRc together with its autoprocessing activity, exciting questions might be raised regarding how the protease exert its activity - whether bound to the outer membrane or in a soluble form. Actually, since the autolytic activity seems to be a requisite for the appearance of activity, it is most likely that APRc is active in a soluble form after its release from the surface of rickettsial cells by an ectodomain shedding-like process. Still, the consistently detection of a band with 21 kDa implies that a significant population of APRc is membrane-embedded. Therefore, we cannot exclude that APRc might also be active against other substrates while still attached to the membrane. Further studies are thus required to elucidate whether APRc is active with the two catalytic domains bound to the outer membrane via their transmembrane domains, or with only one catalytic domain being membrane embedded while the other one is soluble (heterodimer).

Together with our results confirming protease expression and accumulation into the outer membrane in *R. conorii* and *R. rickettsii*, the evidence for the up-regulation of APRc gene expression in *R. prowazekii* Rp22[160] strongly support a potential relevant role of this highly conserved protease in rickettsial pathogenesis. In fact, since virulence determinants are often either secreted to the bacterial cell surface or released into the external environment[309], the membrane-embedded nature of APRc points towards its potential involvement in critical pathogenic mechanisms, such as the modulation of activity/virulence of other rickettsial membrane-localized proteins, including the recently identified Sca family of outer membrane proteins, some of which important virulence factors[182,284]. A comparative genomic analysis of rickettsiae revealed that five of these outer membrane proteins, namely rOmpA (Sca0), Sca1, Sca2, Sca4 (geneD), and rOmpB (Sca5) are highly conserved among the majority of SFG rickettsial species and some of them share homology with the superfamily of proteins produced by pathogenic gram-negative bacteria called autotransporters[184–188,217,284]. As previously mentioned, these proteins are typically composed by 3 domains: a leader sequence that mediates transport across the cell membrane, a passenger sequence that harbors a virulence function (passenger domain), and a transporter sequence that is inserted as a β-barrel into the outer envelope to transport the passenger sequence to the outer surface of the cell wall (autotransporter domain)[310,311]. The Sca4 is an exception to this general structural organization, since it has no β-barrel and thus it is not an autotransporter protein[182]. Amongst Sca family proteins, rOmpB is the most abundant surface protein in *Rickettsia* spp. and was previously described as capable to trigger bacteria internalization in the absence of other virulence factors[186]. Whereas this protein appears to interact with multiple eukaryotic plasma membrane proteins, the bacterial entry is dependent on the interaction between the rOmpB passenger domain and the mammalian Ku70 surface protein[186]. Importantly, *R. rickettsii* Iowa has been reported to be defective in the processing of rOmpB, which is thought to contribute to the avirulence of this strain[142,187]. However, no experimental report has determined so far whether the inability of this rickettsial strain to lyse Vero cells and cause infection in guinea pigs is the result of defective rOmpB processing, some other mutation, or a combination of these two factors. Nevertheless, in contrast with other autotransporter proteins from gram-negative bacteria with auto-proteolytic activity such as SPATEs[310], rOmpB processing is thought to implicate a protease as previous expression studies in *E. coli* have failed to demonstrate autocatalytic activity[186,284]. In view of this**,** it is reasonable to speculate that the protease responsible for this cleavage must be highly conserved among rickettsiae and also membrane-associated. Therefore, we have started addressing this hypothesis and we showed that APRc is indeed sufficient to catalyze the processing of rOmpB *in vitro* and that the generated product is consistent with the cleavage between the passenger and the β-peptide regions. We further confirmed this cleavage by the ability of APRc to cleave a synthetic peptide corresponding to the sequence between both domains of *R. conorii* rOmpB, and which have resulted in the identification of different cleavage products. This result may reflect again the importance of the sequence context/substrate conformation for APRc cleavage specificity. Synthetic peptide substrates quite often do not adopt the same native structure of the substrate and consequently display different conformational constraints. In addition, many proteases are also thought to require non-active site interaction surfaces, or exosites, to recognize and cleave physiological substrates with high specificity and catalytic efficiency. For all these reasons, it is not unexpected that PeprOmpB might be more readily accommodated in the substrate-binding groove of APRc and that different binding positions may be tolerated in light of the broad specificity found *in vitro* for APRc. In addition, the method employed in this study (mass spectrometry) to identify the N-terminal sequence of the cleavage products has remarkable advantages over the Edman sequencing. Despite the fact that the traditional Edman technique is very robust and provides *de novo* capabilities, the sensitivity is in the range of 2-5 pmol of a purified peptide. In contrast, mass spectrometry has routinely been used with peptides in the range of 100 fmol or even less. Taking this into account, we cannot exclude also that the identification of only one cleavage site for *R. typhi* and *R. prowazekii* rOmpB might be attributed to the lower sensitivity of Edman sequencing, suggesting that more than one cleavage may indeed occur *in vivo*. However, our results do not allow the identification of the preferential cleavage site of APRc or of the most abundant cleavage product, for which a quantitative analysis would be required. Nevertheless, considering the pattern of specificity of APRc obtained with PICS, the preferred cleavage sites would be Ala/Val, Ala/Ala, Gly/Asp and Ala/Gly. In fact, although it prefers aromatic amino acids at P1 and P1' position, APRc also tolerates small or positively charged amino acids (see Figure 23), and therefore, these cleavage sites are also in good agreement with the observed specificity preferences for APRc. Furthermore, the diminished capacity of APRc to cleave PeprOmpB under the presence of indinavir and nelfinavir is of major importance, not just for the validation of cleavage specificity, but also to support the use of HIV-1 PR inhibitors to explore how the inhibition of APRc might contribute for a compromised capacity of *Rickettsia* to adhere and invade their target cells. Moreover, attending to the capacity of APRc to process rOmpA *in vitro*, it will be also important to evaluate the proteolytic activity of this protease towards a synthetic peptide corresponding to the sequence between the passenger and the autotransporter domains of rOmpA, in order to give further insights regarding the processing site of this Sca protein (only determined for rOmpB). Altogether, our results clearly unveil APRc as the candidate enzyme for the processing of rOmpB and rOmpA, thereby laying the foundations to study its relevance in an *in vivo* context as well as APRc role in the degradation of other rickettsial Sca proteins and/or host proteins.

Chapter V. General Discussion and Conclusions

Chapter V. General Discussion and Conclusions

A number of studies carried out over the last years have culminated with the identification of very distinct bacterial APs, which are distributed in *MEROPS* database[1] between distant families from different clans, due to great differences in sequence motifs (A1 – clan AA; A8 – clan AC; A24 – clan AD; A25 and A31 – clan AE; A26 – clan AF; A5 and A36 – unassigned clan) (Table 1, page 5). In fact, until recently, APs of families A1 (pepsin) were assumed to be restricted to eukaryotes and the presence of A2 (retropepsin) members in bacteria remains controversial. Importantly, despite retroviral-type APs have been reported in *Bacillus subtilis* (SpoIIGA)[98] and in *Caulobacter crescentus* (PerP)[99], as previously mentioned in Chapter I, their inclusion as retropepsin-type protease members has not been universally accepted mostly because they lack fundamental enzymatic characterization[16].

Other bacterial APs that possess unusual traits have also been reported. Family A24 is represented by the type IV prepilins peptidase (TFPP), and constitutes a novel family of bilobal aspartic protease. Type IV pilin is a protein found on the surface of *Pseudomonas aeruginosa*, *Neisseria gonorrhoeae* and other gram-negative pathogens, as well as in Archaea[312]. In *Pseudomonas aeruginosa*, this bifunctional enzyme is a key determinant in both type IV pilus biogenesis and extracellular protein secretion given its roles as a leader peptidase and methyl transferase. Important secreted proteins include toxins such as cholera toxin of *V. cholerae*[313] and exotoxin A of *Pseudomonas aeruginosa*[314]. TFPP is responsible for endopeptidic cleavage of leader peptides of precursor proteins with type IV pilin precursors, as well as proteins with homologous leader sequences that are essential components of the general secretion pathway found in a variety of gram-negative pathogens. Following removal of the leader peptides, the same enzyme is responsible for the second post-translational modification that characterizes the type IV pilins and their homologues, namely N-methylation of the newly exposed N-terminal amino acid residue[312]. The TFPPs differ from the majority of APs in that the active site Asp residues are not found in the Asp-Thr/Ser-Gly motif, the optimum pH for *in vitro* activity is near neutral as opposed to pH 2–4, and peptidase activity is not inhibited by pepstatin.

Omptins constitutes another well studied family of aspartic proteases found in bacteria. They are a family of structurally related surface proteases found in pathogenic species of the *Enterobacteriaceae*, comprising OmpT and OmpP of *E. coli*, Pla of *Yersinia pestis*, PgtE of *Salmonella*, Pla endopeptidase A of *Erwinia pyrifoliae*, and SopA of *Shigella flexneri*. The sequences lack the signature sequences of classical protease families and Cys residues, have typical features of a β-barrel fold and are resistant against typical inhibitors of APs[315]. Until recently, omptins formed the S18 family of serine peptidases, which as a group are characterized by the Ser-Asp-His catalytic triad[1]. However, the crystal structure of OmpT[316] has revealed a pair of Asp residues at the catalytic site and, on this basis, the omptins were reclassified as APs (family A26, Clan AF). The omptins share a common structural backbone and have minor sequence variations in their surface-exposed regions, which results in differing specificities and functions in infectious diseases. In line with this, they have been demonstrated to be multifunctional surface proteins: besides proteolytic activity, they may enhance bacterial virulence by nonproteolytic functions or promote bacterial adherence to tissue components and invasion into human cells[315].

In 2005, Carrol and Setlow[317] described a protease, named germination protease (GPR), involved on the degradation of small, acid-soluble spore proteins during germination of spores of *Bacillus* and *Clostridium* species. Due to the lack of amino acid sequence homology of GPR with members of the major protease families, the authors classified this protease as an atypical AP after the identification of Asp127 and Asp193 as the catalytic residues, either by site-directed mutagenesis and structural studies[317], and the enzyme included in a new family A25.

Regardless of these relatively few examples on APs in bacteria, no valuable contribution that could support either of the two evolutionary theories involving pepsins and retropepsins (previously discussed in Chapter I) has been provided until the identification of pepsin homologues in prokaryotes[16]. The recent report on the first prokaryotic pepsin homologues in the genome sequences of several γ-proteobacteria[16] and the further validation of shewasin A as an active enzyme[17], brought new insights about pepsin-like ancestors. These important findings clearly suggest that prokaryotic APs may be the archetype of modern eukaryotic APs, implying that the duplication and fusion events have occurred before the divergence of bacteria and eukaryotes. Our current results on RC1339/APRc further support this hypothesis by providing the first experimental evidence that a gene for a single-lobed AP is indeed present in prokaryotes, coding for an active enzyme with properties resembling those of retropepsins. The presence of single-lobed AP genes in prokaryotes suggests that enzymes such as APRc may actually represent the most ancestral forms of these proteases, whereas retroviral proteases may instead correspond to a derived state. Accordingly, the identification of APRc homologues in other α-proteobacteria strengthens the hypothesis that this protease may have originally evolved in the α-proteobacterial lineage - the one that gave rise to mitochondria - and may have later been transferred from the protomitochondrial genome to the ancestral eukaryotic nuclear genome[97,139]. Consequently, the occurrence of this class of enzymes in retroviruses and retrotransposons is thereby suggested to have resulted from the capture and incorporation of a eukaryotic retropepsin gene at an early stage of eukaryotic evolution.

Given the central role of proteins in living systems as essential mediators of biological functions, their evolution has been the object of intense study in the last decades. It is now clear that the increase in protein repertoire of an organism is manly driven by the duplication of sequences coding for one or more domains, the divergence of these duplicated sequences most frequently due to point mutations, insertions and deletions, and the recombination of genes that results in novel arrangements of domains[318]. Such mechanisms, taken repeatedly in the course of evolution, are effective paths to increased protein complexity and diversification[319,320]. Examination of genome sequences and protein structures show that most proteins are formed by combinations of multiple domains linked together in a single polypeptide chain, which have been correlated with gain of novel or modified protein functions. While the basic domain counterparts were already established to a large extent at the time of the 'last common ancestor', other very successful domains evolved later within the archaea, bacteria, or eukaryotes and have been spread by endosymbiosis or lateral transfer into the other kingdoms [321]. These findings clearly support the view of an increased complexity of APs from a common ancestror to the most complex eukaryotes. In fact, based on recognized evolutionary trends toward reduction in archaea and toward complexity in eukarya[322–324], the identification of a gene coding for a retropepsin-like homologue in archaea by Teixeira[258], provides strong evidences to show that the ancestror of APs was presumably a single lobed and soluble protein. Correspondingly, the existence of homodimeric APs with transmembrane domains and the bilobal soluble pepsin-like enzymes with higher complexity found in eukaryotes (comprising pre-, pro- and plant specific inserts elements, for example) in comparison with bacterial pepsin homologues, strengthen the hypothesis of a divergence of at least two lineages branching from a common ancestror of APs. Apparently, the acquisition and retention of transmembrane domains by membrane embedded APs (e.g., APRc and SASPase), and the duplication accompanied by gene fusion mechanisms that gave rise to pepsin-like enzymes, were likely two independent events that gave rise to the two lineages. Along with these divergence processes, APs from different organisms became so dissimilar that their common origin cannot be detected from their sequences, even though they may still fulfill fundamentally the same function. The results presented in this work clearly corroborate this general view. In fact, while rickettsial APs show a low degree of amino acid sequence similarity with their eukaryotic and virus counterparts their structures diverge much more slowly, providing evidence of common ancestry long after their sequence similarity has decayed. The observed conservation of secondary structure between APRc and other retropepsins further strengthen the evolutionary relationships between these proteases. Therefore, solving the three-dimensional structure of APRc and its homologues will likely allow us to gain deeper understanding of this structural proximity with eukaryotic and viral retropepsins, and will provide fundamental information to address questions regarding evolution of APs.

Pathogenic bacteria have evolved a variety of mechanisms to disable cells of the mammalian immune system and create a niche in which they can multiply and disseminate. As part of the arsenal of bacterial virulence factors, some bacteria make use of impressively efficient proteases that have a wide range of biological functions which can be very subtle and specific to help establishing and maintaining an infection[6,226]. Therefore, proteases from pathogens are now recognized important targets for drug discovery, and as a result, many pharmaceutical companies and academic labs around the world are currently putting major efforts on the development of molecules to selectively block the activity of these enzymes without harming normal cellular function[58]. Actually, the role of proteases in bacterial pathogenesis has gained special focus on research over the past few decades since they have been shown not only to degrade structural proteins of the host and cause massive tissue damage, but also to indirectly influence - activate or inhibit - protease cascades of the human body, such as the innate and acquired immune defenses of infected mammals. A familiar example of a bacterial protease known to interact with their hosts during a pathogenic infection is the called lethal factor of Anthrax toxin, a metalloprotease of the pathogenic bacterium *Bacillus anthracis* that specifically cleaves and inactivates MAP kinase kinases[325]. Another example is the botulinum neurotoxin from *Clostridium botulinium*, which hold a metalloprotease domain capable of blocking acetylcholine release at peripheral nerve ending by the cleavage the SNAP-25 protein that plays a role in the storage and depletion of acetylcholine[326]. Botulinum neurotoxin is considered as one of most lethal toxins in nature with a LD_{50} of roughly 0.005–0.05 µg/kg[327].

Owing to the importance of proteases in pathogenesis, future research is expected to show whether proteases may become targets for novel treatment strategies of bacterial infectious diseases of humans, animals and plants. Interestingly, while serine-, cysteine-, and metalloproteases are widely spread in many pathogenic bacteria, much less is known about the role of APs. In fact, besides the aforementioned type IV prepilins[328] and omptin peptidases[329], no other APs (particularly of the retropepsin-type) have been reported so far to participate in bacterial pathogenesis. In this work, taking into account the unique biochemical and enzymatic features of APRc: i) the apparent non-stringent sequence requirement; ii) outer membrane localization and extracellular orientation of recombinant APRc catalytic domain and

iii) autolytic activity suggesting that the soluble biological unit may be released from the surface of rickettsial cells by an ectodomain shedding-like process, we anticipate a potential multi-functional role for this protease on the various stages of rickettsial infection (Figure 32).

The human body has a complex set of overlapping defenses to prevent most of the bacteria it encounters from causing injury, which can be divided in specific and non-specific defenses. The latter ones include antibacterial substances such as complement, phagocytic cells, and the washing action of fluids such as saliva and urine, whereas the specific defenses are cells producing antibodies upon stimulation, and cytotoxic cells. At the critical early period of infection, the non-specific defenses are the host's only defenses and thus it is not surprising that the ability of certain types of bacteria to cause infection depend on characteristics that allow them to evade the primary defense mechanisms of the body. Such characteristics include the so called secreted virulence factors that manipulate or even destroy defense lines of the host. Like other proteases, APRc might play an important role at the various stages of rickettsial infectious process by directly affecting the immunological defense functions through the inactivation of the components of the host immune system, such as circulating antibodies or complement proteins as well as antimicrobial cationic peptides in epithelial or phagocytic cells (Figure 32A)[227,330].

Rickettsial entry into host cells occurs mainly by induced phagocytosis in nonphagocytic cells, a mechanism mediated by different proteins displayed at the surface of the cell[183,185,331]. As we anticipate in this work by the ability of APRc to perform the *in vitro* cleavage of recombinant *R. conorii* rOmpB and rOmpA, this protease may likely be implicated on the degradation and/or maturation of other rickettsial proteins, in particular those located at the outer membrane. Therefore, it will be also important to extend these studies to other Sca proteins such as Sca1 and Sca2, also reported to suffer a proteolytic cleavage for maturation (Figure 32B)[182,188].

Because most intracellular pathogens are frequently taken into the cell via an endocytic or a phagocytic vacuole, they have at least a transient association with a vacuole. However, in contrast with a specialized group of bacteria (*Chlamydia*, *Salmonella*, *Brucella*, *Legionella*, *Coxellia*, and *Mycobacterium*) which reside and replicate within specific vacuoles[332], most common intracellular bacterial pathogens enter cells via endocytosis, followed by rapid escape to the cytoplasm to avoid the lysosomal pathway. As discussed in Chapter I, once within the host cell phagosomal vacuole, phospholipase D and tlyC have been recognized as the major effectors of rickettsial phagosomal escape due to their membranolytic activities. Nevertheless, the acidic phagosomal micro-environment suggests that APRc may also intervene in the formation of vacuole gaps, by degrading membrane proteins of these organelles. Supporting this idea, the secreted IgA1 protease of pathogenic *Neisseria gonorrhoeae* is a serine protease that was found to cleave phagosomal molecules, such as the human lamp-1 membrane protein[333]. By interfering with human lysosomal/phagosomal membrane proteins, IgA1 protease is thought to facilitate the destruction of the phagosomal membrane and the subsequent release of *N. gonorrhoeae* to the host cytosol (Figure 32C)[333].

Once free in the host cytosol, *Rickettsia* requires many components for growth and replication. As detailed by Andersson et al.[139], amino acid metabolism is deficient in *Rickettsia*, and therefore, many amino acids which are likely not synthetized by *Rickettsia* must be provided by the host cell. Since bacterial proteases are assumed to control and modify the environment according to the needs of the bacterium within the host tissue, APRc might contributes for the degradation of host tissues for supplying bacteria with amino acids, similar to that described for other extracellular proteases secreted by many pathogens (Figure 32D)[226].

The increased microvascular permeability resulting from the disruption of adherens junctions between infected endothelial cells with consequent development of inter-endothelial gaps, formation of stress fibers, and conversion of the shape of endothelial cells from polygons to large spindles), is considered the major pathophysiological effect of rickettsial infections[204]. Even though the molecular mechanisms underlying the appearance of these gaps are still poorly understood, they have been suggested to include endothelial cell production of toxic reactive oxygen species, damage to the cell membrane upon rickettsial exit, and cytotoxic T lymphocyte-induced apoptosis of infected endothelial cells[204]. Nevertheless, despite proteases have not been suggested to participate in this process, we cannot exclude that APRc may also directly damage host structures, such as fibrin clots or extracellular matrices, thereby promoting the spread of the infection and dissemination of bacteria across tissue barriers (Figure 32E).

Overall, this hypothesized ability of APRc to perform more than one function or additional catalytic side activities is strongly supported by the report of protein moonlighting/multitasking as a widespread phenomenon in bacterial pathogens[334]. In particular, it has been shown that several proteins from pathogens and other host-associated bacteria with contracting genomes, acquire new or alternate functions, apparently to compensate for gene loss[335]. Accordingly, the evolution of multitasking has been hypothesized to arise from shifts in the selective pressures of remaining genes to favor an increased protein functional diversity[335,336].

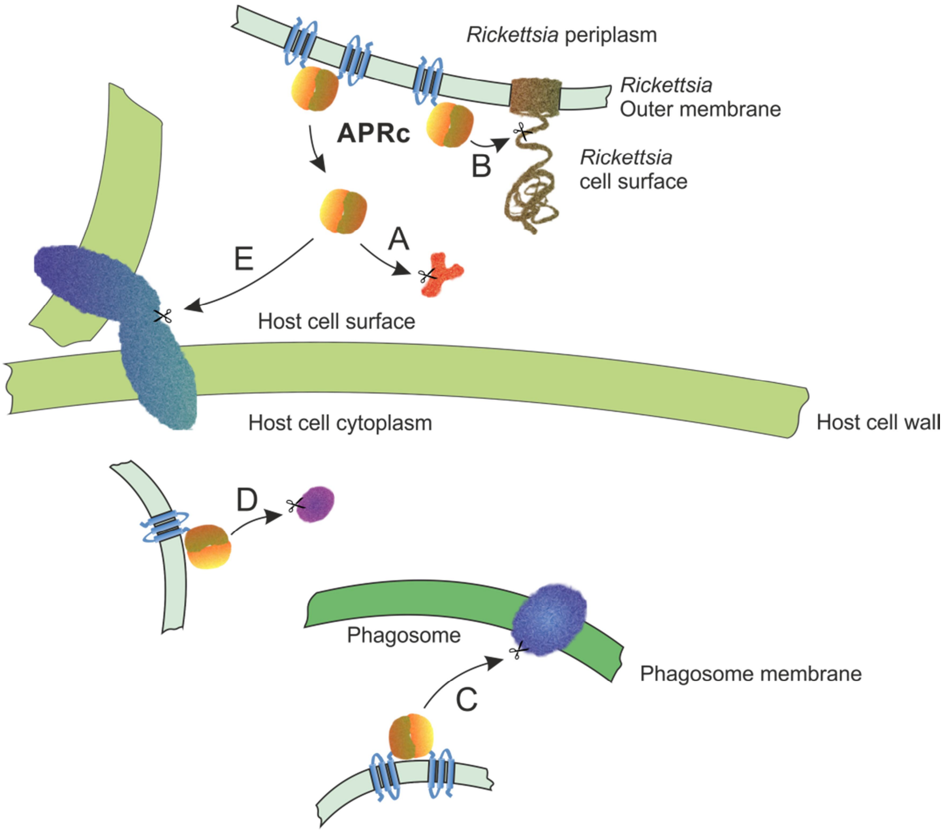


Figure 32. *Schematic representation of proposed APRc biological roles.* The unique biochemical and enzymatic features of APRc suggests a multi-functional role for this protease on the various stages of rickettsial infection. (**A**) One of the proposed functions is the potential implication of APRc on the inactivation of the host immune system components (e.g., antibodies or complement proteins). (**B**) As anticipated in this work, APRc might also be involved in the maturation of other outer membrane proteins of *Rickettsia*, such as the Sca proteins (e.g., rOmpB and rOmpA). (**C**) To facilitate the *Rickettsia* escape from the lysosomal pathway, APRc might cleave human lysosomal/phagosomal membrane proteins leading to the destruction of the phagosomal membrane. (**D**) Inside the host cell cytoplasm, APRc can contribute for the *Rickettsia* supply of amino acids by degrading host proteins. (**E**) The last proposed function of APRc relies on its potential role on the degradation of adherens junction proteins between infected endothelial cells, thereby facilitating the rickettsial infection dissemination. Given its autoprocessing activity, APRc is suggested to exert its function in a soluble form (exemplified by **A** and **E**) or attached to the outer membrane through only one (exemplified by **B** and **D**) or the two catalytic domains (exemplified by **C**).

In the future, it will be critical to validate some of the present findings *in vivo* and study the functional role and the mechanism by which this enzyme can contribute to rickettsial pathogenesis, thereby exploring in more detail the potential of APRc as candidate target for therapeutic inhibition in the treatment of rickettsioses.

Mutagenesis of particular genes is one of the most powerful means to understand how bacteria and their hosts interact during the course of an infection. This approach can be applied to identify and characterize virulence-associated genes[337]. However, although important progresses have been made during past decade, mutagenesis studies have been historically difficult to implement in intracellular bacteria like *Rickettsia*, and genetic manipulation for the development of reliable methods of targeted gene disruption in rickettsiae is still a remaining challenge[338]. So far, three main systems have been successfully used for *Rickettsia* genetic manipulation: homologous recombination, transposition and site-directed mutagenesis. Despite transformants were unstable and difficult to obtain by homologous recombination[339], this was the first report on the direct genetic manipulation of rickettsiae and provided the basis for further work in this field. Transposon mutagenesis using *mariner* element *Himar1* has enabled the insertion and complementation of defective genes with restoration of wild-type phenotype as well as the generation of random gene knockouts[340,341]. Regardless of these advances, a robust method of targeted gene inactivation in rickettsiae is a remaining obstacle, with only a single publication by Driskell and colleagues[195] reporting the use of site-directed mutagenesis for the analysis of a putative rickettsial virulence gene. In this study, the authors targeted the *R. prowazekii pld* (RP819) gene, which encodes a protein with homology to the phospholipase D (PLD) family, by site-directed knockout mutagenesis using homologous recombination. Although quite challenging, the generation of *APRc* mutant in *R. conorii* by site-directed mutagenesis with further phenotypic evaluation, would definitely give a broad understanding on the mechanism by which APRc might contribute to rickettsial pathogenesis and its relevance as a therapeutic target. For instance, the generation of an APRc mutant would offer an elegant way to study the substrate repertoire of this enzyme by one of the two most powerful protein-centric strategies currently used for the identification of protease cleavages sites and substrates: combined fractional diagonal chromatography (COFRADIC)[342] and terminal amine isotopic labeling of substrates (TAILS)[343]. COFRADIC and TAILS are the only N-terminomics approaches that provide both broad coverage and isotopic quantification that is essential for the study of protease's substrate degradome with unknown or broad cleavage-site recognition motifs[343]. For a complete picture of the proteolytic pathways in which APRc might be involved, it would be also interesting to compare the APRc degradome at different time points of infection or even at different points of *Rickettsia* life cycle.

In conclusion, this work provides clear evidences that this new AP from *Rickettsia* is indeed an active enzyme with features resembling those of retropepsin family. The native expression and the outer membrane-embedded nature anticipate a key role on rickettsial virulence through the degradation/maturation of other outer membrane proteins such as rOmpB and rOmpA. All together, the results from this study provide insights into the function of this novel core rickettsial protein and open new lines of research in the study of this complex and intriguing intracellular organism. Additionally, with this work we expect to contribute to start changing the currently accepted evolutionary paradigm of APs, by positioning what we denominate as “prokaryopepsins” as the new archetypes of modern APs.

References

References

1. **Rawlings ND, Waller M, Barrett AJ, Bateman A**. 2014. *MEROPS: The Database of Proteolytic Enzymes, Their Substrates and Inhibitors.* Nucleic Acids Res. **42**:D503–D509.
2. **Vandeputte-Rutten L**. 2002. *Novel Proteases: Common Themes and Surprising Features*. Curr. Opin. Struct. Biol. **12**:704–708.
3. **Sajid M, McKerrow JH**. 2002. *Cysteine Proteases of Parasitic Organisms*. Mol. Biochem. Parasitol. **120**:1–21.
4. **López-Otín C, Matrisian LM**. 2007. *Emerging Roles of Proteases in Tumour Suppression.* Nat. Rev. Cancer **7**:800–808.
5. **Tsiatsiani L, Gevaert K, Van Breusegem F**. 2012. *Natural Substrates of Plant Proteases: How Can Protease Degradomics Extend Our Knowledge?* Physiol. Plant. **145**:28–40.
6. **Ingmer H, Brøndsted L**. 2009. *Proteases in Bacterial Pathogenesis*. Res. Microbiol. **160**:704–710.
7. **Cole SL, Vassar R**. 2008. *The Role of Amyloid Precursor Protein Processing by BACE1, the Beta-Secretase, in Alzheimer Disease Pathophysiology*. J. Biol. Chem. **283**:29621–29625.
8. **López-Otín C, Bond JS**. 2008. *Proteases: Multifunctional Enzymes in Life and Disease.* J. Biol. Chem. **283**:30433–30437.
9. **Castro HC, Abreu PA, Geraldo RB, Loureiro I V, Martins CA, Rodrigues CR**. 2011. *Looking at the Proteases from a Simple Perspective*. J. Mol. Recognit. **24**:165–181.
10. **Van der Hoorn RAL**. 2008. *Plant Proteases: From Phenotypes to Molecular Mechanisms.* Annu. Rev. Plant Biol. **59**:191–223.
11. **Adamson CS**. 2012. *Protease-Mediated Maturation of HIV: Inhibitors of Protease and the Maturation Process*. Mol. Biol. Int. **2012**:1–13.
12. **Saeki K, Ozaki K, Kobayashi T, Ito S**. 2007. *Detergent Alkaline Proteases: Enzymatic Properties, Genes, and Crystal Structures.* J. Biosci. Bioeng. **103**:501–508.
13. **Li Q, Yi L, Marek P, Iverson BL**. 2013. *Commercial Proteases: Present and Future.* FEBS Lett. **587**:1155–1163.
14. **Sumantha A, Larroche C, Pandey A**. 2006. *Microbiology and Industrial Biotechnology of Food-Grade Proteases: A Perspective*. Food Technol. Biotechnol. **44**:211–220.
15. **Schechter I, Berger A**. 1967. *On the Size of the Active Site in Proteases. I. Papain*. Biochem. Biophys. Res. Commun. **27**:157–162.
16. **Rawlings ND, Bateman A**. 2009. *Pepsin Homologues in Bacteria.* BMC Genomics **10**:437–447.
17. **Simões I, Faro R, Bur D, Kay J, Faro C**. 2011. *Shewasin A, an Active Pepsin Homolog from the Bacterium Shewanella Amazonensis.* FEBS J. **278**:3177–86.
18. **Fruton JS**. 2002. *A History of Pepsin and Related Enzymes.* Q. Rev. Biol. **77**:127–147.
19. **Bela Szecsi P, Harboe M**. 2013. *Chapter 5 – Chymosin*, p. 37–42. *In* Handbook of Proteolytic Enzymes. Elsevier.
20. **Fusek M, Mares M, Vetvicka V**. 2013. *Chapter 8 – Cathepsin D*, p. 54–63. *In* Rawlings, ND, Salvesen, G (eds.), Handbook of Proteolytic Enzymes, 1st ed. Elsevier, 1.
21. **Phillips MI, Schmidt-Ott KM**. 1999. *The Discovery of Renin 100 Years Ago.* News Physiol. Sci. **14**:271–274.
22. **Vassar R**. 2001. *The Β-Secretase, BACE*. J. Mol. Neurosci. **17**:157–170.
23. **Tyndall JD a, Nall T, Fairlie DP, Madala PK**. 2010. *Update 1 of : Proteases Universally Recognize Beta Strands In Their Active Sites*. Chem. Rev. **110**:1–31.
24. **Dash C, Kulkarni A, Dunn B, Rao M**. 2003. *Aspartic Peptidase Inhibitors: Implications in Drug Development*. Crit. Rev. Biochem. Mol. Biol. **38**:89–119.
25. **Simões I, Faro C**. 2004. *Structure and Function of Plant Aspartic Proteinases.* Eur. J. Biochem. **271**:2067–20675.

26. **Ge X, Dietrich C, Matsuno M, Li G, Berg H, Xia Y**. 2005. *An Arabidopsis Aspartic Protease Functions as an Anti-Cell-Death Component in Reproduction and Embryogenesis.* EMBO Rep. **6**:282–288.

27. **Dunn BM, Goodenow MM, Gustchina A, Wlodawer A**. 2002. *Retroviral Proteases*. Genome Biol. **3**:3006.1–3006.7.

28. **Goldfarb NE, Dunn BM**. 2013. *Chapter 44 – Human Immunodeficiency Virus 1 Retropepsin*, p. 190–199. *In* Handbook of Proteolytic Enzymes. Elsevier.

29. **Tao N, Liu D, Tang J, Sepulveda P, Marciniszyn J, Chen KC, Huang WY, Lanier JP**. 1973. *Amino-Acid Sequence of Porcine Pepsin*. Proc. Natl. Acad. Sci. U. S. A. **70**:3437–3439.

30. **James MNG, Sielecki AR**. 1983. *Structure and Refinement of Penicillopepsin at 1.8 Å Resolution*. J. Mol. Biol. **163**:299–361.

31. **Cooper JB, Khan G, Taylor G, Tickle IJ, Blundell TL**. 1990. *X-Ray Analyses of Aspartic Proteinases. II. Three-Dimensional Structure of the Hexagonal Crystal Form of Porcine Pepsin at 2.3 A Resolution.* J. Mol. Biol. **214**:199–222.

32. **Navia MA, Fitzgerald PM, McKeever BM, Leu CT, Heimbach JC, Herber WK, Sigal IS, Darke PL, Springer JP**. 1989. *Three-Dimensional Structure of Aspartyl Protease from Human Immunodeficiency Virus HIV-1.* Nature **337**:615–620.

33. **Lapatto R, Blundell T, Hemmings A, Overington J, Wilderspin A, Wood S, Merson JR, Whittle PJ, Danley DE, Geoghegan KF**. 1989. *X-Ray Analysis of HIV-1 Proteinase at 2.7 A Resolution Confirms Structural Homology among Retroviral Enzymes.* Nature **342**:299–302.

34. **Wlodawer A, Miller M, Jaskólski M, Sathyanarayana BK, Baldwin E, Weber IT, Selk LM, Clawson L, Schneider J, Kent SB**. 1989. *Conserved Folding in Retroviral Proteases: Crystal Structure of a Synthetic HIV-1 Protease.* Science **245**:616–621.

35. **Miller M, Jaskólski M, Rao JK, Leis J, Wlodawer A**. 1989. *Crystal Structure of a Retroviral Protease Proves Relationship to Aspartic Protease Family.* Nature **337**:576–579.

36. **Dunn BM**. 2002. *Structure and Mechanism of the Pepsin-like Family of Aspartic Peptidases.* Chem. Rev. **102**:4431–4458.

37. **Cascella M, Micheletti C, Rothlisberger U, Carloni P**. 2005. *Evolutionarily Conserved Functional Mechanics across Pepsin-like and Retroviral Aspartic Proteases.* J. Am. Chem. Soc. **127**:3734–3742.

38. **Andreeva N, Rumsh L**. 2001. *Analysis of Crystal Structures of Aspartic Proteinases: On the Role of Amino Acid Residues Adjacent to the Catalytic Site of Pepsin-like Enzymes*. Protein Sci. 2439–2450.

39. **Barrett A, Woessner J, Rawlings N**. 2013. *Handbook of Proteolytic Enzymes*.

40. **Friedman R, Caflisch A**. 2010. *On the Orientation of the Catalytic Dyad in Aspartic Proteases.* Proteins **78**:1575–1582.

41. **Wlodawer A, Gustchina A**. 2000. *Structural and Biochemical Studies of Retroviral Proteases.* Biochim. Biophys. Acta **1477**:16–34.

42. **Tang J, James MN, Hsu IN, Jenkins JA, Blundell TL**. 1978. *Structural Evidence for Gene Duplication in the Evolution of the Acid Proteases.* Nature **271**:618–621.

43. **Levy Y, Caflisch A**. 2003. *Flexibility of Monomeric and Dimeric HIV-1 Protease*. J. Phys. Chem. B **107**:3068–3079.

44. **Levy Y, Caflisch A, Onuchic JN, Wolynes PG**. 2004. *The Folding and Dimerization of HIV-1 Protease: Evidence for a Stable Monomer from Simulations.* J. Mol. Biol. **340**:67–79.

45. **Tóth G, Borics A**. 2006. *Closing of the Flaps of HIV-1 Protease Induced by Substrate Binding: A Model of a Flap Closing Mechanism in Retroviral Aspartic Proteases*. Biochemistry **45**:6606–6614.

46. **Heaslet H, Rosenfeld R, Giffin M, Lin YC, Tam K, Torbett BE, Elder JH, McRee DE, Stout CD**. 2007. *Conformational Flexibility in the Flap Domains of Ligand-Free HIV Protease.* Acta Crystallogr. D. Biol. Crystallogr. **63**:866–875.

47. **Hornak V, Simmerling C**. 2007. *Targeting Structural Flexibility in HIV-1 Protease Inhibitor Binding.* Drug Discov. Today **12**:132–138.
48. **Ishima R, Torchia DA, Shannon M, Gronenborn AM, John M, Lynch SM, Louis JM**. 2003. *Solution Structure of the Mature HIV-1 Protease Monomer: Insight into the Tertiary Fold and Stability of a Precursor*. J. Biol. Chem. **278**:43311–43319.
49. **Ingr M, Uhlíková T, Stříšovský K, Majerová E, Konvalinka J**. 2003. *Kinetics of the Dimerization of Retroviral Proteases: The "fireman's Grip" and Dimerization*. Protein Sci. **12**:2173–2182.
50. **Strisovsky K, Tessmer U, Langner J, Konvalinka J, Kräusslich HG**. 2000. *Systematic Mutational Analysis of the Active-Site Threonine of HIV-1 Proteinase: Rethinking the "Fireman's Grip" Hypothesis.* Protein Sci. **9**:1631–1641.
51. **Prabu-Jeyabalan M, Nalivaika E, Schiffer CA**. 2002. *Substrate Shape Determines Specificity of Recognition for HIV-1 Protease: Analysis of Crystal Structures of Six Substrate Complexes.* Structure **10**:3693–81.
52. **Ishima R, Ghirlando R, Gronenborn AM, Torchia DA, Louis JM, Tözsér J**. 2001. *Folded Monomer of HIV-1 Protease*. J. Biol. Chem. **276**:49110–49116.
53. **Dunn B**. 1997. *Splitting Image*. Nat. Struct. Mol. Biol. **4**:969–972.
54. **Neurath H**. 1986. *The Versatility of Proteolytic Enzymes.* J. Cell. Biochem. **32**:35–49.
55. **Kageyama T**. 2002. *Pepsinogens, Progastricsins, and Prochymosins: Structure, Function, Evolution, and Development*. Cell. Mol. Life Sci. C. **59**:288–306.
56. **Richter C, Tanaka T, Yada RY**. 1998. *Mechanism of Activation of the Gastric Aspartic Proteinases: Pepsinogen, Progastricsin and Prochymosin*. Biochem. J **490**:481–490.
57. **Lin X, Koelsch G, Loy J, Tang J**. 1995. *Rearranging the Domains of Pepsinogen*. Protein Sci. **4**:159–66.
58. **Mannhold R, Kubinyi H, Folkers G, Ghosh A**. 2010. *Aspartic Acid Proteases as Therapeutic Targets*. WILEY-VCH Verlag & Co. KGaA.
59. **Brynda J, Fábry M, Tichý PJ, Horejsí M, Sedlácek J**. 1995. *Processing, Purification, and Kinetic Characterization of the Gag-Pol Encoded Retroviral Proteinase of Myeloblastosis Associated Virus Expressed in E. Coli.* Adv. Exp. Med. Biol. **362**:485–488.
60. **Yoshinaka Y, Katoh I, Copeland TD, Oroszlan S**. 1985. *Murine Leukemia Virus Protease Is Encoded by the Gag-Pol Gene and Is Synthesized through Suppression of an Amber Termination Codon.* Proc. Natl. Acad. Sci. U. S. A. **82**:1618–16122.
61. **Parkin NT, Chamorro M, Varmus HE**. 1992. *Human Immunodeficiency Virus Type 1 Gag-Pol Frameshifting Is Dependent on Downstream mRNA Secondary Structure: Demonstration by Expression in Vivo.* J. Virol. **66**:5147–5151.
62. **Löchelt M, Flügel RM**. 1996. *The Human Foamy Virus Pol Gene Is Expressed as a Pro-Pol Polyprotein and Not as a Gag-Pol Fusion Protein.* J. Virol. **70**:1033–1040.
63. **Pettit S, Everitt L, Choudhury S, Dunn BM, Kaplan AH**. 2004. *Initial Cleavage of the Human Immunodeficiency Virus Type 1 GagPol Precursor by Its Activated Protease Occurs by an Intramolecular Mechanism*. J. Virol. **78**:8477–8485.
64. **Louis JM, Ishima R, Torchia DA, Weber IT**. 2007. *HIV-1 Protease: Structure, Dynamics, and Inhibition.* Adv. Pharmacol. **55**:261–298.
65. **Louis JM, Aniana A, Weber IT, Sayer JM**. 2011. *Inhibition of Autoprocessing of Natural Variants and Multidrug Resistant Mutant Precursors of HIV-1 Protease by Clinical Inhibitors.* Proc. Natl. Acad. Sci. U. S. A. **108**:9072–9077.
66. **Tözsér J**. 2010. *Comparative Studies on Retroviral Proteases: Substrate Specificity.* Viruses **2**:147–165.
67. **Pettit SC, Lindquist JN, Kaplan AH, Swanstrom R**. 2005. *Processing Sites in the Human Immunodeficiency Virus Type 1 (HIV-1) Gag-Pro-Pol Precursor Are Cleaved by the Viral Protease at Different Rates*. Retrovirology **2**:66–72.
68. **Deu E, Verdoes M, Bogyo M**. 2012. *New Approaches for Dissecting Protease Functions to Improve Probe Development and Drug Discovery.* Nat. Struct. Mol. Biol. **19**:9–16.

69. **Northrop DB**. 2001. *Follow the Protons: A Low-Barrier Hydrogen Bond Unifies the Mechanisms of the Aspartic Proteases.* Acc. Chem. Res. **34**:790–797.

70. **James MN, Sielecki AR, Hayakawa K, Gelb MH**. 1992. *Crystallographic Analysis of Transition State Mimics Bound to Penicillopepsin: Difluorostatine- and Difluorostatone-Containing Peptides.* Biochemistry **31**:3872–3886.

71. **Veerapandian B, Cooper JB, Sali A, Blundell TL, Rosati RL, Dominy BW, Damon DB, Hoover DJ**. 1992. *Direct Observation by X-Ray Analysis of the Tetrahedral "intermediate" of Aspartic Proteinases*. Protein Sci. **1**:322–328.

72. **Davies D**. 1990. *The Structure and Function of the Aspartic Proteinases*. Annu. Rev. Biophys. Biophys. Chem. **19**:189–215.

73. **Brik A, Wong C**. 2003. *HIV-1 Protease: Mechanism and Drug Discovery*. Org. Biomol. Chem. **1**:5–14.

74. **Palashoff M**. 2008. *Determining the Specificity of Pepsin for Proteolytic Digestion*. Northeastern University.

75. **Wang W, Liang TC**. 1994. *Substrate Specificity of Porcine Renin: P1', P1, and P3 Residues of Renin Substrates Are Crucial for Activity.* Biochemistry **33**:14636–14641.

76. **Brier S, Maria G, Carginale V, Capasso A, Wu Y, Taylor RM, Borotto NB, Capasso C, Engen JR**. 2007. *Purification and Characterization of Pepsins A1 and A2 from the Antarctic Rock Cod Trematomus Bernacchii.* FEBS J. **274**:6152–6166.

77. **Hamuro Y, Coales S, Molnar K, Tuske S, Morrow J**. 2008. *Specificity of Immobilized Porcine Pepsin in H/D Exchange Compatible Conditions*. Rapid Commun. Mass Spectrom. **22**:1041–1046.

78. **Paschalidou K, Neumann U, Gerhartz B, Tzougraki C**. 2004. *Highly Sensitive Intramolecularly Quenched Fluorogenic Substrates for Renin Based on the Combination of L-2-Amino-3-(7-Methoxy-4-Coumaryl)propionic Acid with 2,4-Dinitrophenyl Groups at Various Positions.* Biochem. J. **382**:1031–1048.

79. **Nakagawa T, Akaki J, Satou R, Takaya M, Iwata H, Katsurada A, Nishiuchi K, Ohmura Y, Suzuki F, Nakamura Y**. 2007. *The His-Pro-Phe Motif of Angiotensinogen Is a Crucial Determinant of the Substrate Specificity of Renin.* Biol. Chem. **388**:237–246.

80. **Tözsér J, Oroszlan S**. 2003. *Proteolytic Events of HIV-1 Replication as Targets for Therapeutic Intervention*. Curr. Pharm. Des. 1803–1815.

81. **Tözsér J, Bagossi P, Weber IT, Louis JM, Copeland TD, Oroszlan S**. 1997. *Studies on the Symmetry and Sequence Context Dependence of the HIV-1 Proteinase Specificity.* J. Biol. Chem. **272**:16807–16814.

82. **Boross P, Bagossi P, Copeland TD, Oroszlan S, Louis JM, Tözsér J**. 1999. *Effect of Substrate Residues on the P2' Preference of Retroviral Proteinases.* Eur. J. Biochem. **264**:921–9.

83. **Wlodawer A, Vondrasek J**. 1998. *Inhibitors of HIV-1 Protease: A Major Success of Structure-Assisted Drug Design.* Annu. Rev. Biophys. Biomol. Struct. **27**:249–284.

84. **Tomasselli AG, Heinrikson RL**. 1994. *Specificity of Retroviral Proteases: An Analysis of Viral and Nonviral Protein Substrates*, p. 279–301. *In* Methods in Enzymology. Elsevier.

85. **Umezawa H, Aoyagi T, Morishima H, Matsuzaki M, Hamada M**. 1970. *Pepstatin, a New Pepsin Inhibitor Produced by Actinomycetes.* J. Antibiot. (Tokyo). **23**:259–362.

86. **Chitpinityol S, Crabbe MJC**. 1998. *Chymosin and Aspartic Proteinases*. Food Chem. **61**:395–418.

87. **Pearl LH, Taylor WR**. 1987. *A Structural Model for the Retroviral Proteases*. Nature **329**:351–354.

88. **Pokorná J, Machala L, Řezáčová P, Konvalinka J**. 2009. *Current and Novel Inhibitors of HIV Protease*. Viruses **1**:1209–1239.

89. **Velazquez-campoy A, Luque I, Todd MJ, Milutinovich M, Kiso Y, Freire E**. 2000. *Thermodynamic Dissection of the Binding Energetics of KNI-272, a Potent HIV-1 Protease Inhibitor*. Protein Sci. **9**:1801–1809.

90. **Slee DH, Laslo KL, Elder JH, Ollmann IR, Gustchina A, Kervinen J, Zdanov A, Wlodawer A, Wong C-H**. 1995. *Selectivity in the Inhibition of HIV and FIV Protease: Inhibitory and Mechanistic Studies of Pyrrolidine-Containing Alpha-Keto Amide and Hydroxyethylamine Core Structures*. J. Am. Chem. Soc. **117**:11867–11878.
91. **Louis JM, Weber IT, Tözsér J, Clore GM, Gronenborn AM**. 2000. *HIV-1 Protease: Maturation, Enzyme Specificity, and Drug Resistance.* Adv. Pharmacol. **49**:111–146.
92. **Kozal M**. 2004. *Cross-Resistance Patterns among HIV Protease Inhibitors.* AIDS Patient Care STDS **18**:199–208.
93. **Jensen C, Herold P, Brunner HR**. 2008. *Aliskiren: The First Renin Inhibitor for Clinical Treatment.* Nat. Rev. Drug Discov. **7**:399–410.
94. **Rao JK, Erickson JW, Wlodawer A**. 1991. *Structural and Evolutionary Relationships between Retroviral and Eucaryotic Aspartic Proteinases.* Biochemistry **30**:4663–71.
95. **Lin XL, Lin YZ, Koelsch G, Gustchina A, Wlodawer A, Tang J**. 1992. *Enzymic Activities of Two-Chain Pepsinogen, Two-Chain Pepsin, and the Amino-Terminal Lobe of Pepsinogen.* J. Biol. Chem. **267**:17257–17263.
96. **Tang J, Wong RN**. 1987. *Evolution in the Structure and Function of Aspartic Proteases*. J. Cell. Biochem. **33**:53–63.
97. **Krylov DM, Koonin E V**. 2001. *Correspondence: A Novel Family of Predicted Retroviral-like Aspartyl Proteases with a Possible Key Role in Eukaryotic Cell Cycle Control*. Curr. Biol. **11**:584–587.
98. **Imamura D, Zhou R, Feig M, Kroos L**. 2008. *Evidence That the Bacillus Subtilis SpoIIGA Protein Is a Novel Type of Signal-Transducing Aspartic Protease.* J. Biol. Chem. **283**:15287–15299.
99. **Chen JC, Hottes AK, Mcadams HH, Mcgrath PT, Viollier PH, Shapiro L**. 2006. *Cytokinesis Signals Truncation of the PodJ Polarity Factor by a Cell Cycle-Regulated Protease.* EMBO J. **25**:377–386.
100. **Bernard D, Méhul B, Thomas-Collignon A, Delattre C, Donovan M, Schmidt R**. 2005. *Identification and Characterization of a Novel Retroviral-like Aspartic Protease Specifically Expressed in Human Epidermis*. J. Invest. Dermatol. **125**:278–287.
101. **Caputo E, Manco G, Mandrich L, Guardiola J**. 2000. *A Novel Aspartyl Proteinase from Apocrine Epithelia and Breast Tumors.* J. Biol. Chem. **275**:7935–7941.
102. **Sirkis R, Gerst JE, Fass D**. 2006. *Ddi1, a Eukaryotic Protein with the Retroviral Protease Fold.* J. Mol. Biol. **364**:376–387.
103. **Perteguer MJ, Gómez-puertas P, Cañavate C, Dagger F, Gárate T, Valdivieso E**. 2013. *Ddi1-like Protein from Leishmania Major Is an Active Aspartyl Proteinase.* Cell Stress Chaperones **18**:171–181.
104. **Matsui T, Kinoshita-Ida Y, Hayashi-Kisumi F, Hata M, Matsubara K, Chiba M, Katahira-Tayama S, Morita K, Miyachi Y, Tsukita S**. 2006. *Mouse Homologue of Skin-Specific Retroviral-like Aspartic Protease Involved in Wrinkle Formation.* J. Biol. Chem. **281**:27512–27525.
105. **Matsui T, Miyamoto K, Kubo A, Kawasaki H, Ebihara T, Hata K, Tanahashi S, Ichinose S, Imoto I, Inazawa J, Kudoh J, Amagai M**. 2011. *SASPase Regulates Stratum Corneum Hydration through Profilaggrin-to-Filaggrin Processing*. EMBO Mol. Med. **3**:1–14.
106. **Li M, DiMaio F, Zhou D, Gustchina A, Lubkowski J, Dauter Z, Baker D, Wlodawer A**. 2011. *Crystal Structure of XMRV Protease Differs from the Structures of Other Retropepsins*. Nat. Struct. Mol. Biol. **18**:227–229.
107. **Nicholson WL, Allen KE, McQuiston JH, Breitschwerdt EB, Little SE**. 2010. *The Increasing Recognition of Rickettsial Pathogens in Dogs and People.* Trends Parasitol. **26**:205–212.
108. **Walker DH, Ismail N**. 2008. *Emerging and Re-Emerging Rickettsioses: Endothelial Cell Infection and Early Disease Events.* Nat. Rev. Microbiol. **6**:375–386.

109. **Parola P, Paddock CD, Raoult D**. 2005. *Tick-Borne Rickettsioses around the World: Emerging Diseases Challenging Old Concepts*. Society **18**:719–756.
110. **Walker DH**. 2007. *Rickettsiae and Rickettsial Infections: The Current State of Knowledge.* Clin. Infect. Dis. **45 Suppl 1**:S39–44.
111. **Azad AF**. 2007. *Pathogenic Rickettsiae as Bioterrorism Agents*. Clin. Infect. Dis. **45 Suppl 1**:S52–5.
112. **Bechah Y, Capo C, Mege J, Raoult D**. 2008. *Rickettsial Diseases: From Rickettsia-Arthropod Relationships to Pathophysiology and Animal Models*. Future Microbiol. **3**:1–14.
113. **Perlman SJS, Hunter MS, Zchori-Fein E**. 2006. *The Emerging Diversity of Rickettsia*. Proc. R. Soc. **273**:2097–2106.
114. **Gimenez DF**. 1964. *Staining Rickettsiae in Yolk-Sac Cultures*. Stain Technol. **39**:135–40.
115. **Ricketts HT**. 1909. *A Micro-Organism Which Apparently Has a Specific Relationship to Rocky Mountain Spotted Fever*. JAMA J. Am. Med. Assoc. **II**:379–380.
116. **Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, Daszak P**. 2008. *Global Trends in Emerging Infectious Diseases.* Nature **451**:990–993.
117. **Hackstadt T**. 1996. *The Biology of Rickettsiae.* Infect. Agents Dis. **5**:127–143.
118. **Raoult D, Woodward T, Dumler JS**. 2004. *The History of Epidemic Typhus.* Infect. Dis. Clin. North Am. **18**:127–140.
119. **Raoult D, Roux V**. 1999. *The Body Louse as a Vector of Reemerging Human Diseases.* Clin. Infect. Dis. **29**:888–911.
120. **Black FL**. 1959. *Viral and Rickettsial Infections of Man*. Yale J. Biol. Med. **31**:430.
121. **Rocha Lima H**. 1916. *Zur Aetiologie Des Fleckfiebers*. Berliner Klin. Wochenschrift **53**:567–569.
122. **CDC**. 2013. *Centers for Disease Control and Prevention*.
123. **Bechah Y, Capo C, Mege J-L, Raoult D**. 2008. *Epidemic Typhus.* Lancet Infect. Dis. **8**:417–426.
124. **Dumler JS, Walker DH**. 2005. *Rocky Mountain Spotted Fever -Changing Ecology and Persisting Virulence.* N. Engl. J. Med. **353**:551–553.
125. **Cunningham A**. 2005. *A Walk on the Wild Side -Emerging Wildlife Diseases.* BMJ **331**:1214–5.
126. **Blancou J, Chomel B, Belotto A, Meslin F**. 2005. *Emerging or Re-Emerging Bacterial Zoonoses: Factors of Emergence, Surveillance and Control.* Vet. Res. **36**:507–522.
127. **Parola P, Raoult D**. 2006. *Tropical Rickettsioses.* Clin. Dermatol. **24**:191–200.
128. **Rovery C, Brouqui P, Raoult D**. 2008. *Questions on Mediterranean Spotted Fever a Century after Its Discovery.* Emerg. Infect. Dis. **14**:1360–1367.
129. **Sousa R, Nóbrega S**. 2003. *Mediterranean Spotted Fever in Portugal*. Ann. N. Y. Acad. Sci. **990**:285–294.
130. **Parola P**. 2004. *Tick-Borne Rickettsial Diseases: Emerging Risks in Europe.* Comp. Immunol. Microbiol. Infect. Dis. **27**:297–304.
131. **Brouqui P, Parola P, Fournier PE, Raoult D**. 2007. *Spotted Fever Rickettsioses in Southern and Eastern Europe.* FEMS Immunol. Med. Microbiol. **49**:2–12.
132. **Rolain J, Franc M, Davoust B, Raoult D**. 2003. *Molecular Detection of Bartonella Quintana, B. Koehlerae, B. Henselae, B. Clarridgeiae, Rickettsia Felis, and Wolbachia Pipientis in Cat Fleas, France*. Emerg. Infect. Dis. **9**:0–4.
133. **Lindblom A, Severinson K, Nilsson K**. 2010. *Rickettsia Felis Infection in Sweden: Report of Two Cases with Subacute Meningitis and Review of the Literature.* Scand. J. Infect. Dis. **42**:906–909.
134. **Paddock CD**. 2009. *The Science and Fiction of Emerging Rickettsioses*. Ann. N. Y. Acad. Sci. **1166**:133–143.

135. **Verhoeven GS, Alexeeva S, Dogterom M, den Blaauwen T**. 2009. *Differential Bacterial Surface Display of Peptides by the Transmembrane Domain of OmpA.* PLoS One **4**:e6739.

136. **Darby AC, Cho N, Fuxelius H, Westberg J, Andersson SGE**. 2007. *Intracellular Pathogens Go Extreme: Genome Evolution in the Rickettsiales.* Trends Genet. **23**:511–520.

137. **Finlay BB, Falkow S**. 1997. *Common Themes in Microbial Pathogenicity Revisited.* Microbiol. Mol. Biol. Rev. **61**:136–169.

138. **Olsen GJ, Woese CR, Overbeek R**. 1994. *The Winds of (evolutionary) Change: Breathing New Life into Microbiology.* J. Bacteriol. **176**:1–6.

139. **Andersson SGES, Zomorodipour A, Andersson JO, Sicheritz-ponte T, Alsmark UCM, Podowski RM, Naslund AK, Winkler HH, Eriksson A, Kurland CG**. 1998. *The Genome Sequence of Rickettsia Prowazekii and the Origin of Mitochondria*. Nature **396**:133–140.

140. **Pierlé SA, Dark MJ, Dahmen D, Palmer GH, Brayton KA**. 2012. *Comparative Genomics and Transcriptomics of Trait-Gene Association.* BMC Genomics **13**:669–676.

141. **Weisburg WG, Dobson ME, Samuel JE, Dasch GA, Mallavia LP, Baca O, Mandelco L, Sechrest JE, Weiss E, Woese CR**. 1989. *Phylogenetic Diversity of the Rickettsiae.* J. Bacteriol. **171**:4202–4206.

142. **Merhej V, Raoult D**. 2011. *Rickettsial Evolution in the Light of Comparative Genomics*. Biol. Rev. Camb. Philos. Soc. **86**:379–405.

143. **Gillespie JJ, Williams K, Shukla M, Snyder EE, Nordberg EK, Shane M, Dharmanolla C, Rainey D, Soneja J, Shallom JM, Dongre N, Ceraul SM, Vishnubhat ND, Wattam R, Purkayastha A, Czar M, Crasta O, Setubal JC, Azad AF, Sobral BS**. 2008. *Rickettsia Phylogenomics: Unwinding the Intricacies of Obligate Intracellular Life.* PLoS One **3**:e2018.

144. **Gillespie, J. J., Nordberg, E., Azad, A.F., Sobral BW**. 2012. *Phylogeny and Comparative Genomics: The Shifting Landscape in the Genomics Era*, p. 84–141. *In* A. F. Azad, Palmer, GH (ed.), Intracellular Pathogens II: Rickettsiales. American Society of Microbiology.

145. **Dumler JS, Barbet AF, Bekker CPJ, Dasch GA, Palmer GH, Ray SC, Rikihisa Y, Rurangirwa FR**. 2001. *Reorganization of Genera in the Families Rickettsiaceae and Anaplasmataceae in the Order Rickettsiales: Unification of Some Species of Ehrlichia with Anaplasma, Cowdria with Ehrlichia and Ehrlichia with Neorickettsia, Descriptions of Six New Species Combi*. Int. J. Syst. Evol. Microbiol. **51** :2145–2165.

146. **Roux V, Raoult D**. 1995. *Phylogenetic Analysis of the Genus Rickettsia by 16S rDNA Sequencing.* Res. Microbiol. **146**:385–396.

147. **Fournier PPP, Dumler JS, Greub G, Zhang J, Wu Y, Raoult D**. 2003. *Gene Sequence-Based Criteria for Identification of New Rickettsia Isolates and Description of Rickettsia Heilongjiangensis Sp. Nov.* J. Clin. Microbiol. **41**:5456–5465.

148. **Raoult D, Roux V**. 1997. *Rickettsioses as Paradigms of New or Emerging Infectious Diseases.* Clin. Microbiol. Rev. **10**.

149. **Ogata H, La Scola B, Audic S, Renesto P, Blanc G, Robert C, Fournier P, Claverie J, Raoult D**. 2006. *Genome Sequence of Rickettsia Bellii Illuminates the Role of Amoebae in Gene Exchanges between Intracellular Pathogens*. PLoS Genet. **2**:e76.

150. **Gillespie JJ, Beier MS, Rahman MS, Ammerman NC, Shallom JM, Purkayastha A, Sobral BS, Azad AF**. 2007. *Plasmids and Rickettsial Evolution: Insight from Rickettsia Felis.* PLoS One **2**:e266.

151. **Hagimori T, Abe Y, Date S, Miura K**. 2006. *The First Finding of a Rickettsia Bacterium Associated with Parthenogenesis Induction among Insects.* Curr. Microbiol. **52**:97–101.

152. **Reeves WK, Loftis AD, Szumlas DE, Abbassy MM, Helmy IM, Hanafi HA, Dasch GA**. 2007. *Rickettsial Pathogens in the Tropical Rat Mite Ornithonyssus Bacoti (Acari: Macronyssidae) from Egyptian Rats (Rattus Spp.).* Exp. Appl. Acarol. **41**:101–107.

153. **Weinert LA, Werren JH, Aebi A, Stone GN, Jiggins FM**. 2009. *Evolution and Diversity of Rickettsia Bacteria*. BMC Biol. **15**:1–15.

154. **Vitorino L, Chelo IMI, Bacellar F, Zé-Zé L**. 2007. *Rickettsiae Phylogeny: A Multigenic Approach*. Microbiology **153**:160–168.

155. **Zhu Y, Fournier P, Eremeeva M, Raoult D**. 2005. *Proposal to Create Subspecies of Rickettsia Conorii Based on Multi-Locus Sequence Typing and an Emended Description of Rickettsia Conorii.* BMC Microbiol. **5**:1–11.

156. **Williams KP, Sobral BW, Dickerman AW**. 2007. *A Robust Species Tree for the Alphaproteobacteria.* J. Bacteriol. **189**:4578–4586.

157. **Emelyanov V V**. 2001. *Rickettsiaceae, Rickettsia-like Endosymbionts, and the Origin of Mitochondria.* Biosci. Rep. **21**:1–17.

158. **Pallen MJ, Wren BW**. 2007. *Bacterial Pathogenomics.* Nature **449**:835–842.

159. **Suhre K, Vestris G, Blanc G, Ogata H, Robert C, Audic S, Claverie J-M, Raoult D**. 2007. *Reductive Genome Evolution from the Mother of Rickettsia.* PLoS Genet. **3**:e14.

160. **Bechah Y, Karkouri K El, Mediannikov O, Leroy Q, Mege J-L, Pelletier N, Robert C, Me C, Raoult D, Médigue C**. 2010. *Genomic, Proteomic, and Transcriptomic Analysis of Virulent and Avirulent Rickettsia Prowazekii Reveals Its Adaptive Mutation Capabilities*. Genome Res. **20**:655–663.

161. **Ogata H, Audic S, Renesto-Audiffren P**. 2001. *Mechanisms of Evolution in Rickettsia Conorii and R. Prowazekii*. Science (80-.). **293**:2093–2098.

162. **Andersson SG, Kurland CG**. 1998. *Reductive Evolution of Resident Genomes*. Trends Microbiol. **6**:263–268.

163. **Wixon J**. 2001. *Featured Organism: Reductive Evolution in Bacteria: Buchnera Sp., Rickettsia Prowazekii and Mycobacterium Leprae*. Comp. Funct. Genomics **2**:44–48.

164. **Lawrence JG, Hendrix RW, Casjens S**. 2001. *Where Are the Pseudogenes in Bacterial Genomes?* Trends Microbiol. **9**:535–540.

165. **Andersson JO, Andersson SG**. 2001. *Pseudogenes, Junk DNA, and the Dynamics of Rickettsia Genomes*. Mol. Biol. Evol. **18**:829–839.

166. **Baldridge G, Burkhardt N, Felsheim R, Kurtti T, Munderloh U**. 2008. *Plasmids of the pRM/pRF Family Occur in Diverse Rickettsia Species*. Appl. Environ. Microbiol. **74**:645–652.

167. **Felsheim RF, Kurtti TJ, Munderloh UG**. 2009. *Genome Sequence of the Endosymbiont Rickettsia Peacockii and Comparison with Virulent Rickettsia Rickettsii: Identification of Virulence Factors.* PLoS One **4**:e8361.

168. **Yu XAM, Walker D**. 2012. *Rickettsia and Rickettsial Diseases*, p. 179–192. *In* Morse, SA (ed.), Bioterrorism.

169. **Martinez JJ, Cossart P**. 2004. *Early Signaling Events Involved in the Entry of Rickettsia Conorii into Mammalian Cells.* J. Cell Sci. **117**:5097–5106.

170. **Teysseire N, Boudier J a, Raoult D**. 1995. *Rickettsia Conorii Entry into Vero Cells*. Infect. Immun. **63**:366–374.

171. **Walker DH, Valbuena GA, Olano JP**. 2003. *Pathogenic Mechanisms of Diseases Caused by Rickettsia*. Ann. N. Y. Acad. Sci. **990**:1–11.

172. **Azad A, Beard C**. 1998. *Rickettsial Pathogens and Their Arthropod Vectors.* Emerg. Infect. Dis. **4**:179–186.

173. **Parola P, Raoult D**. 2001. *Ticks and Tickborne Bacterial Diseases in Humans: An Emerging Infectious Threat*. Clin. Infect. Dis. **32**:897–928.

174. **Beard CB, Durvasula R V, Richards FF**. 1998. *Bacterial Symbiosis in Arthropods and the Control of Disease Transmission.* Emerg. Infect. Dis. **4**:581–591.

175. **Burgdorfer W, Brinton L**. 1975. *Mechanisms of Transovarial Infection of Spotted Fever Rickettsiae in Ticks*. Ann. N. Y. Acad. Sci. **266**:61–72.

176. **Macaluso KRK, Sonenshine DDE, Ceraul SM, Azad AF**. 2002. *Rickettsial Infection in Dermacentor Variabilis (Acari: Ixodidae) Inhibits Transovarial Transmission of a Second Rickettsia*. J. Med. Entomol. **39**:809–813.

177. **Reháček J**. 1984. *Rickettsia Slovaca, the Organism and Its Ecology*. Academia.

178. **Eremeeva ME, Dasch GA**. 2009. *Closing the Gaps between Genotype and Phenotype in Rickettsia Rickettsii.* Ann. N. Y. Acad. Sci. **1166**:12–26.
179. **Zemtsova G, Killmaster LF, Mumcuoglu KY, Levin ML**. 2010. *Co-Feeding as a Route for Transmission of Rickettsia Conorii Israelensis between Rhipicephalus Sanguineus Ticks.* Exp. Appl. Acarol. **52**:383–92.
180. **Gillespie J, Ammerman N, Beier MS, Sobral BS, Azad AF**. 2009. *Louse-and Flea-Borne Rickettsioses: Biological and Genomic Analyses*. Vet. Res. **40**:1–13.
181. **Heinzen RA, Hayes SF, Peacock MG, Hackstadt T**. 1993. *Directional Actin Polymerization Associated with Spotted Fever Group Rickettsia Infection of Vero Cells.* Infect. Immun. **61**:1926–35.
182. **Blanc G, Ngwamidiba M, Ogata H, Fournier P, Claverie J, Raoult D**. 2005. *Molecular Evolution of Rickettsia Surface Antigens: Evidence of Positive Selection*. Mol. Biol. Evol. **22**:2073–2083.
183. **Hillman R, Baktash Y, Martinez JJ**. 2013. *OmpA-Mediated Rickettsial Adherence to and Invasion of Human Endothelial Cells Is Dependent upon Interaction with α2β1 Integrin*. Cell. Microbiol. **15**:727–741.
184. **Riley SP, Goh KC, Hermanas TM, Cardwell MM, Chan YGY, Martinez JJ**. 2010. *The Rickettsia Conorii Autotransporter Protein Sca1 Promotes Adherence to Nonphagocytic Mammalian Cells.* Infect. Immun. **78**:1895–1904.
185. **Cardwell MM, Martinez JJ**. 2009. *The Sca2 Autotransporter Protein from Rickettsia Conorii Is Sufficient to Mediate Adherence to and Invasion of Cultured Mammalian Cells.* Infect. Immun. **77**:5272–5280.
186. **Chan YGY, Cardwell MM, Hermanas TM, Uchiyama T, Martinez JJ**. 2009. *Rickettsial Outer-Membrane Protein B (rOmpB) Mediates Bacterial Invasion through Ku70 in an Actin, c-Cbl, Clathrin and Caveolin 2-Dependent Manner.* Cell. Microbiol. **11**:629–644.
187. **Hackstadt T, Messer R, Cieplak W, Peacock MG**. 1992. *Evidence for Proteolytic Cleavage of the 120-Kilodalton Outer Membrane Protein of Rickettsiae: Identification of an Avirulent Mutant Deficient in Processing.* Infect. Immun. **60**:159–165.
188. **Chan YG, Riley SP, Martinez JJ**. 2010. *Adherence to and Invasion of Host Cells by Spotted Fever Group Rickettsia Species.* Front. Microbiol. **1**:139.
189. **UCHIYAMA T**. 2003. *Adherence to and Invasion of Vero Cells by Recombinant Escherichia Coli Expressing the Outer Membrane Protein rOmpB of Rickettsia Japonica*. Ann. N. Y. Acad. Sci. **990**:585–590.
190. **Veiga E, Matsuyama S, Martinez JJ, Seveau S, Cossart P**. 2005. *Ku70, a Component of DNA-Dependent Protein Kinase, Is a Mammalian Receptor for Rickettsia Conorii.* Cell **123**:1013–1023.
191. **Renesto P, Samson L, Ogata H, Azza S, Fourquet P, Gorvel J-P, Heinzen R a, Raoult D**. 2006. *Identification of Two Putative Rickettsial Adhesins by Proteomic Analysis*. Res. Microbiol. **157**:605–612.
192. **Gouin E, Gantelet H, Egile C, Lasa I, Ohayon H, Villiers V, Gounon P, Sansonetti PJ, Cossart P**. 1999. *A Comparative Study of the Actin-Based Motilities of the Pathogenic Bacteria Listeria Monocytogenes, Shigella Flexneri and Rickettsia Conorii.* J. Cell Sci. **112**:1697–1708.
193. **Whitworth T, Popov V, Yu X**. 2005. *Expression of the Rickettsia Prowazekii Pld or tlyC Gene in Salmonella Enterica Serovar Typhimurium Mediates Phagosomal Escape*. Infect. Immun. **73**:6668–6673.
194. **Renesto P, Dehoux P, Gouin E, Touqui L, Cossart P, Raoult D**. 2003. *Identification and Characterization of a Phospholipase D-Superfamily Gene in Rickettsiae*. J. Infect. Dis. **188**:1276–1283.
195. **Driskell LO, Yu X, Zhang L, Liu Y, Popov VL, Walker DH, Tucker AM, Wood DO**. 2009. *Directed Mutagenesis of the Rickettsia Prowazekii Pld Gene Encoding Phospholipase D.* Infect. Immun. **77**:3244–3248.

196. **Walker DH, Feng HM, Popov VL**. 2001. *Rickettsial Phospholipase A2 as a Pathogenic Mechanism in a Model of Cell Injury by Typhus and Spotted Fever Group Rickettsiae.* Am. J. Trop. Med. Hyg. **65**:936–942.

197. **Rahman MS, Ammerman NC, Sears KT, Ceraul SM, Azad AF**. 2010. *Functional Characterization of a Phospholipase A(2) Homolog from Rickettsia Typhi.* J. Bacteriol. **192**:3294–3303.

198. **Radulovic S, Troyer JM, Beier MS, Lau AOT, Azad AF**. 1999. *Identification and Molecular Analysis of the Gene Encoding Rickettsia Typhi Hemolysin.* Infect. Immun. **67**:6104–6108.

199. **Bhavsar AP, Guttman JA, Finlay BB**. 2007. *Manipulation of Host-Cell Pathways by Bacterial Pathogens.* Nature **449**:827–834.

200. **Jeng RL, Goley ED, D'Alessio JA, Chaga OY, Svitkina TM, Borisy GG, Heinzen R a, Welch MD**. 2004. *A Rickettsia WASP-like Protein Activates the Arp2/3 Complex and Mediates Actin-Based Motility.* Cell. Microbiol. **6**:761–769.

201. **Kleba B, Clark TR, Lutter EI, Ellison DW, Hackstadt T**. 2010. *Disruption of the Rickettsia Rickettsii Sca2 Autotransporter Inhibits Actin-Based Motility.* Infect. Immun. **78**:2240–2247.

202. **Mansueto P, Vitale G, Cascio A, Seidita A, Pepe I, Carroccio A, Rosa S, Rini GB, Cillari E, Walker DH**. 2012. *New Insight into Immunity and Immunopathology of Rickettsial Diseases*. Clin. Dev. Immunol. **2012**:1–26.

203. **Mahajan SK**. 2012. *Rickettsial Diseases.* J. Assoc. Physicians India **60**:37–44.

204. **Valbuena G, Walker DH**. 2005. *Changes in the Adherens Junctions of Human Endothelial Cells Infected with Spotted Fever Group Rickettsiae.* Virchows Arch. an Int. J. Pathol. **446**:379–382.

205. **Davidson MG, Breitschwerdt EB, Walker DH, Levy MG, Carlson CS, Hardie EM, Grindem CA, Nasisse MP**. 1990. *Vascular Permeability and Coagulation during Rickettsia Rickettsii Infection in Dogs.* Am. J. Vet. Res. **51**:165–170.

206. **HARRELL GT, AIKAWA JK**. 1949. *Pathogenesis of Circulatory Failure in Rocky Mountain Spotted Fever; Alterations in the Blood Volume and the Thiocyanate Space at Various Stages of the Disease.* Arch. Intern. Med. **83**:331–347.

207. **Kaabia N, Letaief A**. 2009. *Characterization of Rickettsial Diseases in a Hospital-Based Population in Central Tunisia*. Ann. N. Y. Acad. Sci. **1166**:167–171.

208. **Sahni S, Rydkina E**. 2009. *Host-Cell Interactions with Pathogenic Rickettsia Species*. Future Microbiol. **4**:323–339.

209. **Billings A, Feng H, Olano J, Walker D**. 2001. *Rickettsial Infection in Murine Models Activates an Early Anti-Rickettsial Effect Mediated by NK Cells and Associated with Production of Gamma Interferon.* Am. J. Pathol. **65**:52–56.

210. **Rydkina E, Silverman DJ, Sahni SK**. 2005. *Similarities and Differences in Host Cell Signaling Following Infection with Different Rickettsia Species.* Ann. N. Y. Acad. Sci. **1063**:203–206.

211. **Balraj P, Renesto P, Raoult D**. 2009. *Advances in Rickettsia Pathogenicity.* Ann. N. Y. Acad. Sci. **1166**:94–105.

212. **Walker D, Olano J, Feng H**. 2001. *Critical Role of Cytotoxic T Lymphocytes in Immune Clearance of Rickettsial Infection*. Infect. Immun. **69**:1841–1846.

213. **Sahni SK, Van Antwerp DJ, Eremeeva ME, Silverman DJ, Marder VJ, Sporn LA**. 1998. *Proteasome-Independent Activation of Nuclear Factor kappaB in Cytoplasmic Extracts from Human Endothelial Cells by Rickettsia Rickettsii.* Infect. Immun. **66**:1827–1833.

214. **Joshi SG, Francis CW, Silverman DJ, Sahni SK**. 2003. *Nuclear Factor Kappa B Protects against Host Cell Apoptosis during Rickettsia Rickettsii Infection by Inhibiting Activation of Apical and Effector Caspases and Maintaining Mitochondrial Integrity.* Infect. Immun. **71**:4127–4136.

215. **Sporn L, Sahni S, Lerner N, Marder V, Silverman D, Turpin L, Schwab A**. 1997. *Rickettsia Rickettsii Infection of Cultured Human Endothelial Cells Induces NF-kappaB Activation*. Infect. Immun. **65**:2786–2791.
216. **Anacker RL, Mann RE, Gonzales C**. 1987. *Reactivity of Monoclonal Antibodies to Rickettsia Rickettsii with Spotted Fever and Typhus Group Rickettsiae.* J. Clin. Microbiol. **25**:167–171.
217. **Uchiyama T, Kawano H, Kusuhara Y**. 2006. *The Major Outer Membrane Protein rOmpB of Spotted Fever Group Rickettsiae Functions in the Rickettsial Adherence to and Invasion of Vero Cells.* Microbes Infect. **8**:801–809.
218. **Feng H, Whitworth T, Walker D**. 2004. *Effect of Antibody on the Rickettsia-Host Cell Interaction*. Infect. Immun. **72**:3524–3530.
219. **Paddock CD, Greer PW, Ferebee TL, Singleton J, McKechnie DB, Treadwell TA, Krebs JW, Clarke MJ, Holman RC, Olson JG, Childs JE, Zaki SR**. 1999. *Hidden Mortality Attributable to Rocky Mountain Spotted Fever: Immunohistochemical Detection of Fatal, Serologically Unconfirmed Disease.* J. Infect. Dis. **179**:1469–1476.
220. **Ormsbee R, Peacock M, Philip R, Casper E, Plorde J, Gabre-Kidan T, Wright L**. 1978. *Antigenic Relationships between the Typhus and Spotted Fever Groups of Rickettsiae.* Am. J. Epidemiol. **108**:53–59.
221. **La Scola B, Raoult D**. 1997. *Laboratory Diagnosis of Rickettsioses: Current Approaches to Diagnosis of Old and New Rickettsial Diseases.* J. Clin. Microbiol. **35**:2715–2727.
222. **Renvoisé A, Rolain J-MJ, Socolovschi C, Raoult D**. 2012. *Widespread Use of Real-Time PCR for Rickettsial Diagnosis.* FEMS Immunol. Med. Microbiol. **64**:126–129.
223. **Stenos J, Graves SR, Unsworth NB**. 2005. *A Highly Sensitive and Specific Real-Time PCR Assay for the Detection of Spotted Fever and Typhus Group Rickettsiae.* Am. J. Trop. Med. Hyg. **73**:1083–1085.
224. **Dumler J, Palmer G, Azad A**. 2012. *Clinical Disease: Current Treatment and New Challenges.* Intracell. Pathog. II Rickettsiales 1–39.
225. **Walker D**. 2009. *The Realities of Biodefense Vaccines against Rickettsia*. Vaccine **27**:52–55.
226. **Lebrun I, Marques-Porto R, Pereira AS, Pereira A, Perpetuo EA**. 2009. *Bacterial Toxins: An Overview on Bacterial Proteases and Their Action as Virulence Factors.* Mini Rev. Med. Chem. **9**:820–828.
227. **Władyka B, Pustelny K**. 2008. *Regulation of Bacterial Protease Activity*. Cell. Mol. Biol. Lett. **13**:212–229.
228. **Potempa J, Travis J**. 2000. *Proteinases as Virulence Factors in Bacterial Diseases and as Potential Targets for Therapeutic Intervention with Proteinase Inhibitors*, p. 159–188. *In* Helm, K, Korant, B, Cheronis, J (eds.), Proteases as Targets for Therapy SE - 9. Springer Berlin Heidelberg.
229. **Rahman M, Simser J, Macaluso KR, Azad AF**. 2003. *Molecular and Functional Analysis of the lepB Gene, Encoding a Type I Signal Peptidase from Rickettsia Rickettsii and Rickettsia Typhi*. J. Bacteriol. **185**:4578–4584.
230. **Rahman MS, Ceraul SM, Dreher-lesnick SM, Beier MS, Azad AF**. 2007. *The lspA Gene , Encoding the Type II Signal Peptidase of Rickettsia Typhi: Transcriptional and Functional Analysis*. J. Bacteriol. **189**:336–341.
231. **Temenak JJ, Anderson BE, McDonald GA**. 2001. *Molecular Cloning, Sequence and Characterization of cjsT, a Putative Protease from Rickettsia Rickettsii.* Microb. Pathog. **30**:221–228.
232. **Gillespie JJ, Ammerman NC, Dreher-lesnick SM, Rahman MS, Worley MJ, Setubal JC, Sobral BS, Azad AF, Micah J**. 2009. *An Anomalous Type IV Secretion System in Rickettsia Is Evolutionarily Conserved*. PLoS One **4**:e4833.
233. **Ammerman NC, Gillespie JJ, Neuwald AF, Sobral BW, Azad AF**. 2009. *A Typhus Group-Specific Protease Defies Reductive Evolution in Rickettsiae.* J. Bacteriol. **191**:7609–7613.

234. **Ratner L, Haseltine W, Patarca R, Livak KJ, Starcich B, Josephs SF, Doran ER, Rafalski JA, Whitehorn EA, Baumeister K, Ivanoff L, Petteway SR, Pearson ML, Lautenberger JA, Papas TS, Ghrayeb J, Chang NT, Gallo RC, Wong-Staal F**. 1985. *Complete Nucleotide Sequence of the AIDS Virus, HTLV-III*. Nature **313**:277–284.

235. **Thompson JD, Higgins DG, Gibson TJ**. 1994. *CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice.* Nucleic Acids Res. **22**:4673–4680.

236. **Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJA, Vaughan R**. 2003. *The InterPro Database, 2003 Brings Increased Coverage and New Features*. Nucleic Acids Res. **31**:315–318.

237. **Tusnády GE, Simon I**. 2001. *The HMMTOP Transmembrane Topology Prediction Server.* Bioinformatics **17**:849–850.

238. **Pei J, Kim B-H, Grishin N V**. 2008. *PROMALS3D: A Tool for Multiple Protein Sequence and Structure Alignments.* Nucleic Acids Res. **36**:2295–2300.

239. **Simões I, Faro R, Bur D, Faro C**. 2007. *Characterization of Recombinant CDR1, an Arabidopsis Aspartic Proteinase Involved in Disease Resistance.* J. Biol. Chem. **282**:31358–31365.

240. **Raghavan R, Hicks LD, Minnick MF**. 2008. *Toxic Introns and Parasitic Intein in Coxiella Burnetii: Legacies of a Promiscuous Past*. J. Bacteriol. **190**:5934–5943.

241. **Voth DE, Howe D, Beare PA, Vogel JP, Unsworth N, Samuel JE, Heinzen RA**. 2009. *The Coxiella Burnetii Ankyrin Repeat Domain-Containing Protein Family Is Heterogeneous, with C-Terminal Truncations That Influence Dot/Icm-Mediated Secretion.* J. Bacteriol. **191**:4232–4242.

242. **Fabrini R, De Luca A, Stella L, Mei G, Orioni B, Ciccone S, Federici G, Lo Bello M, Ricci G**. 2009. *Monomer-Dimer Equilibrium in Glutathione Transferases: A Critical Re-Examination.* Biochemistry **48**:10473–10482.

243. **Böhm G, Muhr R, Jaenicke R**. 1992. *Quantitative Analysis of Protein Far UV Circular Dichroism Spectra by Neural Networks*. Protein Eng.

244. **Mahalingam B, Boross P, Wang Y-F, Louis JM, Fischer CC, Tozser J, Harrison RW, Weber IT**. 2002. *Combining Mutations in HIV-1 Protease to Understand Mechanisms of Resistance.* Proteins **48**:107–116.

245. **Sayer JM, Liu F, Ishima R, Weber IT, Louis JM**. 2008. *Effect of the Active Site D25N Mutation on the Structure, Stability, and Ligand Binding of the Mature HIV-1 Protease.* J. Biol. Chem. **283**:13459–13470.

246. **Louis JM, Nashed NT, Parris KD, Kimmel AR, Jerina DM, Nashedti NT, Jerinat DM**. 1994. *Kinetics and Mechanism of Autoprocessing of Human Immunodeficiency Virus Type 1 Protease from an Analog of the Gag-Pol Polyprotein.* Proc. Natl. Acad. Sci. U. S. A. **91**:7970–7974.

247. **Wan M, Takagi M, Loh BN, Xu XZ, Imanaka T**. 1996. *Autoprocessing : An Essential Step for the Activation of HIV-1 Protease*. Biochem. J. **316**:569–573.

248. **Leitner A, Walzthoeni T, Kahraman A**. 2010. *Probing Native Protein Structures by Chemical Cross-Linking, Mass Spectrometry, and Bioinformatics*. Mol. Cell. Proteomics 1–47.

249. **Dinman J**. 2013. *Biophysical Approaches to Translational Control of Gene Expression*. Springer.

250. **Hartl MJ, Wöhrl BM, Rösch P, Schweimer K**. 2008. *The Solution Structure of the Simian Foamy Virus Protease Reveals a Monomeric Protein.* J. Mol. Biol. **381**:141–149.

251. **Hartl MJ, Schweimer K, Reger MH, Schwarzinger S, Bodem J, Osch PR, Rösch P, Wöhrl BM**. 2010. *Formation of Transient Dimers by a Retroviral Protease*. Biochem. J. **427**:197–203.
252. **Fodor SK, Vogt VM**. 2002. *Characterization of the Protease of a Fish Retrovirus, Walleye Dermal Sarcoma Virus*. J. Virol. **76**:4341–4349.
253. **Ido E, Han HP, Kezdy FJ, Tang J**. 1991. *Kinetic Studies of Human Immunodeficiency Virus Type 1 Protease and Its Active-Site Hydrogen Bond Mutant A28S*. J. Biol. Chem. **266**:24359–24366.
254. **Louis JM, Wondrak EM, Kimmel AR, Wingfield PT, Nashed NT**. 1999. *Proteolytic Processing of HIV-1 Protease Precursor, Kinetics and Mechanism*. J. Biol. Chem. **274**:23437–23442.
255. **Tang C, Louis JM, Aniana A, Suh J-Y, Clore GM**. 2008. *Visualizing Transient Events in Amino-Terminal Autoprocessing of HIV-1 Protease.* Nature **455**:693–696.
256. **Lee E-G, Stenbak CR, Linial ML**. 2013. *Foamy Virus Assembly with Emphasis on Pol Encapsidation*. Viruses **5**:886–900.
257. **Louis JM, Ishima R, Nesheiwat I, Pannell LK, Lynch SM, Torchia DA, Gronenborn AM**. 2003. *Revisiting Monomeric HIV-1 Protease. Characterization and Redesign for Improved Properties.* J. Biol. Chem. **278**:6085–6092.
258. **Teixeira P**. 2013. *Bacterial Retropepsin-like Proteases: The Evidence from Legionella Pneumophila*. University of Coimbra.
259. **Bindra JS, Sibanda BL, Blundellt T, Hobart PM, Fogliano M, Dominy BW, Chirgwin JM**. 1984. *Computer Graphics Modelling of Human Renin. Specificity, Catalytic Activity and Intron-Exon Junctions.* FEBS Lett. **174**:102–111.
260. **Yamauchi T, Nagahama M, Hori H, Murakami K**. 1988. *Functional Characterization of Asp-317 Mutant of Human Renin Expressed in COS Cells.* FEBS Lett. **230**:205–208.
261. **Matúz K, Mótyán J, Li M, Wlodawer A, Tőzsér J**. 2012. *Inhibition of XMRV and HIV-1 Proteases by Pepstatin A and Acetyl-Pepstatin*. FEBS J. **279**:3276–3286.
262. **Reymond J-L**. 2006. *Enzyme Assays*. John Wiley & Sons.
263. **Hubbard SJ**. 1998. *The Structural Aspects of Limited Proteolysis of Native Proteins.* Biochim. Biophys. Acta **1382**:191–206.
264. **López-Otín C, Overall CM**. 2002. *Protease Degradomics: A New Challenge for Proteomics.* Nat. Rev. Mol. Cell Biol. **3**:509–519.
265. **Hills R, Mazzarella R, Fok K, Liu M, Nemirovskiy O, Leone J, Zack MD, Arner EC, Viswanathan M, Abujoub A, Muruganandam A, Sexton DJ, Bassill GJ, Sato AK, Malfait A-M, Tortorella MD**. 2007. *Identification of an ADAMTS-4 Cleavage Motif Using Phage Display Leads to the Development of Fluorogenic Peptide Substrates and Reveals Matrilin-3 as a Novel Substrate.* J. Biol. Chem. **282**:11101–11109.
266. **Thornberry NA, Chapman KT, Nicholson DW**. 2000. *Determination of Caspase Specificities Using a Peptide Combinatorial Library.* Methods Enzymol. **322**:100–110.
267. **Turk BE, Huang LL, Piro ET, Cantley LC**. 2001. *Determination of Protease Cleavage Site Motifs Using Mixture-Based Oriented Peptide Libraries.* Nat. Biotechnol. **19**:661–667.
268. **Van den Berg BHJ, Tholey A**. 2012. *Mass Spectrometry-Based Proteomics Strategies for Protease Cleavage Site Identification.* Proteomics **12**:516–529.
269. **Diamond SL**. 2007. *Methods for Mapping Protease Specificity.* Curr. Opin. Chem. Biol. **11**:46–51.
270. **Schilling O, Overall CM**. 2008. *Proteome-Derived, Database-Searchable Peptide Libraries for Identifying Protease Cleavage Sites*. Nat. Biotechnol. **26**:685–695.
271. **Schilling O, Huesgen PF, Barré O, Auf dem Keller U, Overall CM**. 2011. *Characterization of the Prime and Non-Prime Active Site Specificities of Proteases by Proteome-Derived Peptide Libraries and Tandem Mass Spectrometry*. Nat. Protoc. **6**:111–120.
272. **Doucet A, Overall CM**. 2008. *Protease Proteomics: Revealing Protease in Vivo Functions Using Systems Biology Approaches.* Mol. Aspects Med. **29**:339–358.

273. **Wessel D, Flügge UI**. 1984. *A Method for the Quantitative Recovery of Protein in Dilute Solution in the Presence of Detergents and Lipids.* Anal. Biochem. **138**:141–143.

274. **Craig R, Beavis R**. 2004. *TANDEM: Matching Proteins with Tandem Mass Spectra*. Bioinformatics **20**:1466–1467.

275. **Keller A, Nesvizhskii AI, Kolker E, Aebersold R**. 2002. *Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search*. Anal. Chem. **74**:5383–5392.

276. **Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J**. 2003. *TM4: A Free, Open-Source System for Microarray Data Management and Analysis.* Biotechniques **34**:374–378.

277. **Schilling O, auf dem Keller U, Overall CM**. 2011. *Factor Xa Subsite Mapping by Proteome-Derived Peptide Libraries Improved Using WebPICS, a Resource for Proteomic Identification of Cleavage Sites*. Biol. Chem. **392**:1031–1037.

278. **Dunn BM**. 2013. *Chapter 54 – Feline Immunodeficiency Virus Retropepsin*, p. 230–234. *In* Handbook of Proteolytic Enzymes. Elsevier.

279. **Nikaido H**. 1994. *Isolation of Outer Membranes.* Methods Enzymol. **235**:225–34.

280. **Riley SP, Patterson JL, Nava S, Martinez JJ**. 2014. *Pathogenic Rickettsia Species Acquire Vitronectin from Human Serum to Promote Resistance to Complement-Mediated Killing.* Cell. Microbiol. 1–13.

281. **Hobb RI, Fields JA, Burns CM, Thompson SA**. 2009. *Evaluation of Procedures for Outer Membrane Isolation from Campylobacter Jejuni.* Microbiology **155**:979–988.

282. **Tao H, Liu W, Simmons BN, Harris HK, Cox TC, Massiah MA**. 2010. *Purifying Natively Folded Proteins from Inclusion Bodies Using Sarkosyl, Triton X-100, and CHAPS*. Biotechniques **48**:61–64.

283. **Marani P, Wagner S, Baars L, Genevaux P, Gier JDE, Nilsson I, Casadio R, Heijne G Von**. 2006. *New Escherichia Coli Outer Membrane Proteins Identified through Prediction and Experimental Verification*. Protein Sci. **15**:884–889.

284. **Chan Y, Riley S, Chen E, Martinez J**. 2011. *Molecular Basis of Immunity to Rickettsial Infection Conferred through Outer Membrane Protein B*. Infect. Immun. **79**:2303–2313.

285. **Raskin DM, Seshadri R, Pukatzki SU, Mekalanos JJ**. 2006. *Bacterial Genomics and Pathogen Evolution.* Cell **124**:703–714.

286. **Clifton LA, Skoda MWA, Daulton EL, Hughes A V, Le Brun AP, Lakey JH, Holt SA**. 2013. *Asymmetric Phospholipid: Lipopolysaccharide Bilayers; a Gram-Negative Bacterial Outer Membrane Mimic*. J. R. Soc. Interface **10**:1–11.

287. **Elofsson A, von Heijne G**. 2007. *Membrane Protein Structure: Prediction versus Reality.* Annu. Rev. Biochem. **76**:125–140.

288. **Koebnik R, Locher KP, Van Gelder P**. 2000. *Structure and Function of Bacterial Outer Membrane Proteins: Barrels in a Nutshell.* Mol. Microbiol. **37**:239–253.

289. **Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M**. 1977. *The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures.* J. Mol. Biol. **112**:535–542.

290. **Saier MH**. 2006. *Protein Secretion and Membrane Insertion Systems in Gram-Negative Bacteria*. J. Membr. Biol. **1**:414–419.

291. **Ruiz N, Kahne D, Silhavy TJ**. 2006. *Advances in Understanding Bacterial Outer-Membrane Biogenesis*. Nat. Rev. Microbiol. **4**:57–66.

292. **Weiner JH, Li L**. 2008. *Proteome of the Escherichia Coli Envelope and Technological Challenges in Membrane Proteome Analysis.* Biochim. Biophys. Acta **1778**:1698–7113.

293. **Liu J, Rost B**. 2001. *Comparing Function and Structure between Entire Proteomes*. Protein Sci. **10**:1970–1979.

294. **Dong C, Beis K, Nesper J, Brunkan-Lamontagne AL, Clarke BR, Whitfield C, Naismith JH**. 2006. *Wza the Translocon for E. Coli Capsular Polysaccharides Defines a New Class of Membrane Protein.* Nature **444**:226–229.
295. **Ziegler K, Benz R, Schulz GE**. 2008. *A Putative Alpha-Helical Porin from Corynebacterium Glutamicum.* J. Mol. Biol. **379**:482–491.
296. **Jiang J, Tong J, Tan K, Gabriel K**. 2012. *From Evolution to Pathogenesis: The Link between Β-Barrel Assembly Machineries in the Outer Membrane of Mitochondria and Gram-Negative Bacteria*. Int. J. Mol. Sci. **13**:8038–8050.
297. **Dimmer KS, Rapaport D**. 2010. *The Enigmatic Role of Mim1 in Mitochondrial Biogenesis*. Eur. J. Cell Biol. **89**:212–215.
298. **Thornton N, Stroud DA, Milenkovic D, Guiard B, Pfanner N, Becker T**. 2010. *Two Modular Forms of the Mitochondrial Sorting and Assembly Machinery Are Involved in Biogenesis of Alpha-Helical Outer Membrane Proteins.* J. Mol. Biol. **396**:540–549.
299. **Werner J, Misra R**. 2005. *YaeT (Omp85) Affects the Assembly of Lipid-Dependent and Lipid-Independent Outer Membrane Proteins of Escherichia Coli.* Mol. Microbiol. **57**:1450–1459.
300. **Manning DS, Reschke DK, Judd RC**. 1998. *Omp85 Proteins of Neisseria Gonorrhoeae and Neisseria Meningitidis Are Similar to Haemophilus Influenzae D-15-Ag and Pasteurella Multocida Oma87*. Microb. Pathog. **25**:11–21.
301. **Loosmore SM, Yang YP, Coleman DC, Shortreed JM, England DM, Klein MH**. 1997. *Outer Membrane Protein D15 Is Conserved among Haemophilus Influenzae Species and May Represent a Universal Protective Antigen against Invasive Disease.* Infect. Immun. **65**:3701–3707.
302. **Gentle I, Gabriel K, Beech P, Waller R, Lithgow T**. 2004. *The Omp85 Family of Proteins Is Essential for Outer Membrane Biogenesis in Mitochondria and Bacteria.* J. Cell Biol. **164**:19–24.
303. **Paetzel M, Karla A, Strynadka NCJ, Dalbey RE**. 2002. *Signal Peptidases.* Chem. Rev. **102**:4549–4580.
304. **Von Heijne G**. 2006. *Membrane-Protein Topology*. Nat. Rev. Mol. cell Biol. **7**:909–918.
305. **Tusnády GE, Simon I**. 2001. *Topology of Membrane Proteins*. J. Chem. Inf. Comput. Sci. **41**:364–368.
306. **Seppälä S, Slusky JJS, Lloris-Garcerá P, Rapp M, von Heijne G**. 2010. *Control of Membrane Protein Topology by a Single C-Terminal Residue*. Science (80-.). **1698**:1698–700.
307. **Andersson H, Bakker E, von Heijne G**. 1992. *Different Positively Charged Amino Acids Have Similar Effects on the Topology of a Polytopic Transmembrane Protein in Escherichia Coli.* J. Biol. Chem. **267**:1491–1495.
308. **Skach WR**. 2009. *Cellular Mechanisms of Membrane Protein Folding.* Nat. Struct. Mol. Biol. **16**:606–612.
309. **Wilson JW, Schurr MJ, Leblanc CL, Ramamurthy R, Buchanan KL, Nickerson CA**. 2002. *Mechanisms of Bacterial Pathogenicity.* Postgrad. Med. J. **78**:216–224.
310. **Dautin N**. 2010. *Serine Protease Autotransporters of Enterobacteriaceae (SPATEs): Biogenesis and Function.* Toxins (Basel). **2**:1179–1206.
311. **Dautin N, Bernstein H**. 2011. *Residues in a Conserved Α-Helical Segment Are Required for Cleavage but Not Secretion of an Escherichia Coli Serine Protease Autotransporter Passenger Domain*. J. Bacteriol. **193**:3748–3756.
312. **LaPointe CF, Taylor RK**. 2000. *The Type 4 Prepilin Peptidases Comprise a Novel Family of Aspartic Acid Proteases.* J. Biol. Chem. **275**:1502–1510.
313. **Marsh JW, Taylor RK**. 1998. *Identification of the Vibrio Cholerae Type 4 Prepilin Peptidase Required for Cholera Toxin Secretion and Pilus Formation.* Mol. Microbiol. **29**:1481–1492.

314. **Nunn DN, Lory S**. 1992. *Components of the Protein-Excretion Apparatus of Pseudomonas Aeruginosa Are Processed by the Type IV Prepilin Peptidase.* Proc. Natl. Acad. Sci. U. S. A. **89**:47–51.

315. **Haiko J, Suomalainen M, Ojala T, Lähteenmäki K, Korhonen TK**. 2009. *Invited Review: Breaking Barriers Attack on Innate Immune Defences by Omptin Surface Proteases of Enterobacterial Pathogens.* Innate Immun. **15**:67–80.

316. **Vandeputte-Rutten L, Kramer RA, Kroon J, Dekker N, Egmond MR, Gros P**. 2001. *Crystal Structure of the Outer Membrane Protease OmpT from Escherichia Coli Suggests a Novel Catalytic Site.* EMBO J. **20**:5033–5039.

317. **Carroll T, Setlow P**. 2005. *Site-Directed Mutagenesis and Structural Studies Suggest That the Germination Protease, GPR, in Spores of Bacillus Species Is an Atypical Aspartic Acid Protease*. J. Bacteriol. **187**:7119–7125.

318. **Chothia C, Gough J, Vogel C, Teichmann SA**. 2003. *Evolution of the Protein Repertoire.* Science **300**:1701–1703.

319. **Wolfe K, Shields D**. 1997. *Molecular Evidence for an Ancient Duplication of the Entire Yeast Genome*. Nature **387**:708–713.

320. **Söding J, Lupas A**. 2003. *More than the Sum of Their Parts: On the Evolution of Proteins from Peptides.* Bioessays **25**:837–846.

321. **Pál C, Papp B, Lercher MJ**. 2006. *An Integrated View of Protein Evolution*. Nat. Rev. Genet. **7**:337–348.

322. **Forterre P**. 2013. *The Common Ancestor of Archaea and Eukarya Was Not an Archaeon.* Archaea **2013**:1–18.

323. **Boussau B, Blanquart S, Necsulea A, Lartillot N, Gouy M**. 2008. *Parallel Adaptations to High Temperatures in the Archaean Eon.* Nature **456**:942–945.

324. **Groussin M, Gouy M**. 2011. *Adaptation to Environmental Temperature Is a Major Determinant of Molecular Evolutionary Rates in Archaea.* Mol. Biol. Evol. **28**:2661–2674.

325. **Young J a T, Collier RJ**. 2007. *Anthrax Toxin: Receptor Binding, Internalization, Pore Formation, and Translocation.* Annu. Rev. Biochem. **76**:243–265.

326. **Rossetto O, Pirazzini M, Bolognese P, Rigoni M, Montecucco C**. 2011. *An Update on the Mechanism of Action of Tetanus and Botulinum Neurotoxins.* Acta Chim. Slov. **58**:702–707.

327. **Montecucco C, Molgó J**. 2005. *Botulinal Neurotoxins: Revival of an Old Killer*. Curr. Opin. Pharmacol. **5**:274–279.

328. **Sexton JA, Vogel JP**. 2002. *Type IVB Secretion by Intracellular Pathogens.* Traffic **3**:178–185.

329. **Hritonenko V, Stathopoulos C**. 2007. *Omptin Proteins: An Expanding Family of Outer Membrane Proteases in Gram-Negative Enterobacteriaceae*. Mol. Membr. Biol. **24**:395–406.

330. **Blom A, Hallström T, Riesbeck K**. 2009. *Complement Evasion Strategies of Pathogens—acquisition of Inhibitors and beyond*. Mol. Immunol. **46**:2808–2817.

331. **Riley S, Patterson J, Martinez JJ**. 2012. *The Rickettsial OmpB B-Peptide of Rickettsia Conorii Is Sufficient To Facilitate Factor H-Mediated Serum Resistance*. Infect. Immun. **80**:2735–2743.

332. **Huston WM**. 2010. *Bacterial Proteases from the Intracellular Vacuole Niche; Protease Conservation and Adaptation for Pathogenic Advantage*. FEMS Immunol. Med. Microbiol. **59**:1–10.

333. **Hauck CR, Meyer TF**. 1997. *The Lysosomal/phagosomal Membrane Protein H-Lamp-1 Is a Target of the IgA1 Protease of Neisseria Gonorrhoeae*. FEBS Lett. **405**:86–90.

334. **Henderson B, Martin A**. 2011. *Bacterial Virulence in the Moonlight: Multitasking Bacterial Moonlighting Proteins Are Virulence Determinants in Infectious Disease.* Infect. Immun. **79**:3476–3491.

335. **Catrein I, Herrmann R**. 2011. *The Proteome of Mycoplasma Pneumoniae, a Supposedly "Simple" Cell*. Proteomics **11**:3614–3632.

336. **Kelkar Y, Ochman H**. 2013. *Genome Reduction Promotes Increase in Protein Functional Complexity in Bacteria*. Genetics **193**:303–307.

337. **Saenz HL, Dehio C**. 2005. *Signature-Tagged Mutagenesis: Technical Advances in a Negative Selection Method for Virulence Gene Identification.* Curr. Opin. Microbiol. **8**:612–619.

338. **Wood DO**. 2000. *Genetic Manipulation of Rickettsiae: A Preview*. Society **68**:6091–6093.

339. **Rachek LI, Tucker AM, Winkler HH, Wood DO**. 1998. *Transformation of Rickettsia Prowazekii to Rifampin Resistance.* J. Bacteriol. **180**:2118–24.

340. **Felsheim R, Herron M, Nelson CM, Burkhardt NY, Barbet AF, Kurtti TJ, Munderloh UG**. 2006. *Transformation of Anaplasma Phagocytophilum*. BMC Biotechnol. **6**:1–9.

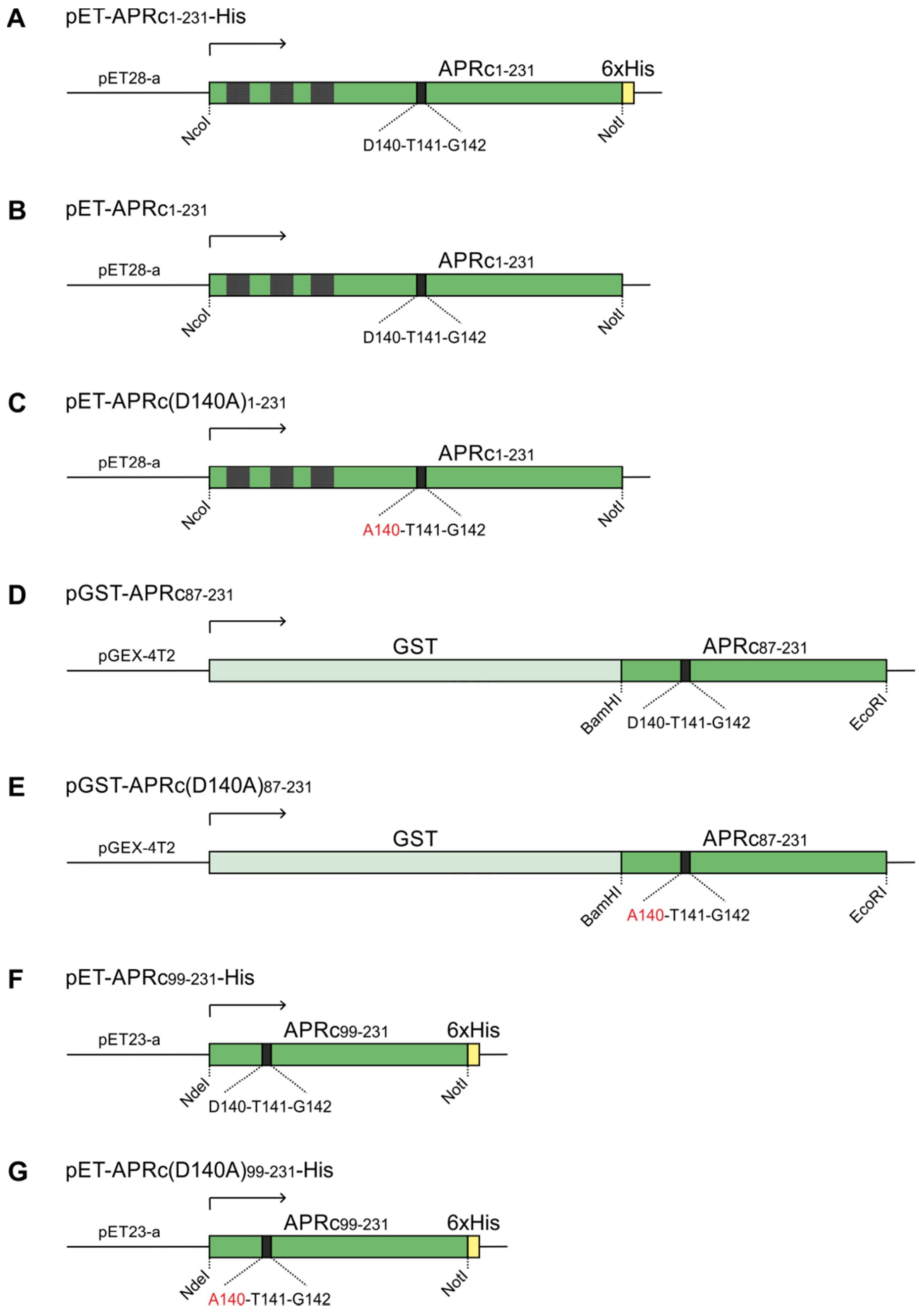
341. **Clark TR, Lackey AM, Kleba B, Driskell LO, Lutter EI, Martens C, Wood DO, Hackstadt T**. 2011. *Transformation Frequency of a Mariner-Based Transposon in Rickettsia Rickettsii*. J. Bacteriol. **193**:4993–4995.

342. **Staes A, Van Damme P, Helsens K, Demol H, Vandekerckhove J, Gevaert K**. 2008. *Improved Recovery of Proteome-Informative, Protein N-Terminal Peptides by Combined Fractional Diagonal Chromatography (COFRADIC).* Proteomics **8**:1362–1370.

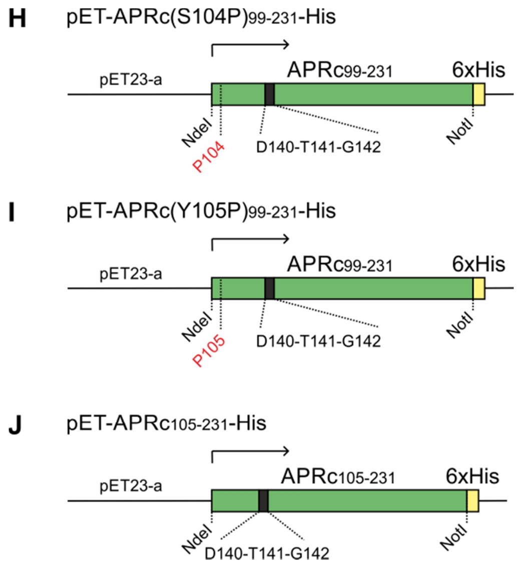
343. **auf dem Keller U, Prudova A, Gioia M, Butler GS, Overall CM**. 2010. *A Statistics-Based Platform for Quantitative N-Terminome Analysis and Identification of Protease Cleavage Products.* Mol. Cell. Proteomics **9**:912–927.

Supplementary Material

Supplementary Material



Supplementary Figure 1. *Schematic representation of the different constructs encoding APRc.* The name of each construct is indicated on the top of each panel, along with name of the expression vector, fusion tag (GST: Glutathione S-transferase tag; 6xHis: hexa His tag) and the restriction sites used. Grey boxes represent the three transmembrane domains and the black box indicates the catalytic active site. Amino acids substitutions are highlighted in red.



Supplementary Figure 1 (cont.). *Schematic representation of the different constructs encoding APRc.* The name of each construct is indicated on the top of each panel, along with name of the expression vector, fusion tag (GST: Glutathione S-transferase tag; 6xHis: hexa His tag) and the restriction sites used. Grey boxes represent the three transmembrane domains and the black box indicates the catalytic active site. Amino acids substitutions are highlighted in red.

Supplementary Table 1. APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem. Peptides identified by LC-MS/MS spectrum-to-sequence assignment with Mascot and X!Tandem are listed with PeptideProphet probability score, calculated neutral mass and one exemplary accession number of a matching Uniprot protein entry is listed. This data was further processed and rendered non-redundant for generation of cleavage specificity profiles.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
AAAALAAAAVK	0.9884	1042.5923	Q8TAQ2
AAAAPAKVEAK	0.9788	1169.6556	Q8NHW5
AAAGAVGSVVGQIAK	0.9460	1413.7728	Q14914
AAAGYDVEKNNSR	0.9726	1509.6960	Q02539
AAAIAYGLDK	0.8889	1107.5712	P11021
AAAIAYGLDK	0.8592	1107.5712	P11021
AAAIAYGLDKK	0.9983	1263.6975	P54652
AAAIAYGLDKK	0.9994	1263.6975	P54652
AAAIAYGLDKK	0.9996	1263.6975	P54652
AAAIAYGLDKK	0.8204	1263.6897	Q91883
AAAIAYGLDKRE	0.9624	1392.7149	P11021
AAALAYGLDKSEDK	0.9811	1594.7991	P38646
AAASIANIVK	0.9051	1072.6029	P17987
AAAVDAGMAMAGQSPVLR	0.8664	1802.8555	P26599
AAEKLQVVGR	0.9845	1185.6618	O43175
AAGAGATHSPPTDLVWK	0.9997	1793.8849	P02545
AAGAGATHSPPTDLVWK	0.9988	1793.8849	P02545
AAGLFLPGSVGITDPCESGNFR	1.0000	2352.0957	Q12905
AAGLFLPGSVGITDPCESGNFR	1.0000	2352.0957	Q12905
AAGLSVPNVHGALAPLAIPSAAAAAAAAGR	0.9955	2723.4619	P26599-2
AAGLSVPNVHGALAPLAIPSAAAAAAAAGR	1.0000	2723.4619	P26599-2
AAGLSVPNVHGALAPLAIPSAAAAAAAAGR	0.9954	2723.4619	P26599-2
AAGSTAGSLR	0.7338	977.4678	Q8TEJ3
AAHVEYSTAAR	0.9925	1262.5792	P49411
AALILVADNAGGSHASK	0.9854	1709.8849	Q7Z5L9
AALKNPPINTK	0.8115	1309.7506	O15511
AAMADTFLEHMCR	0.9958	1639.6693	P14618
AAMADTFLEHMCR	0.8411	1639.6619	P00548
AAMTLLSDASHLPK	0.9239	1569.7973	Q8NBX0
AAPVAAATTAAPAAAAAPAK	0.9801	1777.9474	P05388
AAQLHLQLQSK	0.9932	1351.7360	P54920
AASAPVLAVAGLGDSNQFFR	0.9944	2078.0333	Q9UHL4
AASAPVLAVAGLGDSNQFFR	1.0000	2078.0333	Q9UHL4

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
AASSSSLEK	0.8859	994.4719	P62736
AASVSSSSLV	0.9962	994.4645	Q5VS55
AASVSSSSLV	0.9644	994.4645	Q5VS55
AASVSSSSLV	0.9884	994.4645	Q5VS55
AASVSSSSLV	0.9763	994.4645	Q5VS55
AASVSSSSLV	0.9986	994.4645	Q5VS55
AASVSSSSLV	0.9718	994.4645	Q5VS55
AASVSSSSLV	0.9665	994.4645	Q5VS55
AASVSSSSLV	0.9825	994.4645	Q5VS55
AASVSSSSLV	0.9942	994.4645	Q5VS55
AASVSSSSLV	0.9334	994.4645	Q5VS55
AASVSSSSLV	0.9962	994.4645	Q5VS55
AASVSSSSLV	0.9813	994.4645	Q5VS55
AASVSSSSLV	0.9891	994.4645	Q5VS55
AASVSSSSLV	0.8684	994.4645	Q5VS55
AATLLANHSLR	0.9973	1253.6628	P13489
AAVDTSSEITTK	0.9873	1337.6388	P01252
AAVSNLVR	0.7803	916.4878	P18206
ADALLIIPK	0.9958	1068.6331	Q92526
ADAPMFVMGVNHEK	0.9990	1660.7416	P10096
ADAPMFVMGVNHEK	0.9802	1660.7416	P10096
ADIKAKAQLVK	0.8479	1355.8289	Q76L83
AEGIHTGQFVYCGK	0.9896	1681.7671	P62917
AEHQINLIK	0.9822	1180.6352	P25398
AEILELAGNAAR	0.9858	1314.6680	P04908
AEVLELAGNASK	0.9946	1316.6724	Q71UI9
AFADALLIIPK	0.9928	1286.7386	Q92526
AGGDVCVDR	0.8103	1035.4192	Q96RT8
AGPNTNGSQFFICTAK	0.9998	1827.8362	A2BFH1
AGPTALLAHEIGFGSK	0.9995	1683.8732	Q99497
AGPTALLAHEIGFGSK	0.9947	1683.8732	Q99497
AGPTALLAHEIGFGSK	0.9727	1683.8732	Q99497
AGPTALLAHEIGFGSK	0.9990	1683.8732	Q99497
AGPTALLAHEIGFGSK	0.8201	1683.8732	Q99497
AGPTALLAHEIGFGSK	0.9903	1683.8657	Q5E946
AGPTALLAHEIGFGSK	0.9999	1683.8657	Q5E946
AGPVAEYLK	0.9992	1062.5498	Q01518
AGPVAEYLK	0.9794	1062.5424	Q01518
AGPVAEYLK	0.9444	1062.5424	Q01518

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
AGQCGNQIGAK	0.9914	1218.5563	Q13885
AHAVTQLANR	0.9212	1167.5897	P78371
AHAVTQLANR	0.9974	1167.5897	P78371
AHLDATTVLSR	0.9966	1270.6418	P06576
AHLDATTVLSR	0.9965	1270.6418	P06576
AHLDATTVLSR	0.9976	1270.6418	P06576
AHTFNPK	0.7271	929.4507	P28838
AIAEAWAR	0.9977	974.4722	Q71U36
AIAEAWAR	0.9678	974.4722	Q71U36
AIEHADFAGVER	0.9583	1401.6425	P78371
AIGLSVADLAESIMK	0.7667	1632.8467	P00338
AILDAVGDDIPVQ	0.7746	1412.6861	A9HWC3
AILGMDVLCQAK	0.9994	1433.7159	O00148
AILGMDVLCQAK	0.9979	1433.7159	O00148
AINPELLQLLPLHPK	0.9999	1811.0457	Q99661
AIPSAAAAAAAAGR	0.9997	1255.6421	P26599
AITATQK	0.9843	847.4552	P04406
AITQVLLLANPQK	0.9965	1523.8823	Q9UJY5
ALAASALPALVMSK	0.9495	1457.8064	P36578
ALALFGGEPK	0.8501	1117.5920	P49736
ALCSLHSIGK	0.9912	1200.6073	P14174
ALCSLHSIGK	0.9748	1200.6073	P14174
ALCSLHSIGK	0.9325	1200.6073	P14174
ALCSLHSIGK	0.9430	1200.6073	P14174
ALFEDTNLCAIHAK	0.9998	1717.8172	P02302
ALFPPVEFPAPR	0.9590	1427.7275	P49327
ALFPPVEFPAPR	0.7824	1427.7275	P49327
ALGWVAMAPKPGPYVK	0.9900	1843.9807	Q01518
ALGWVAMAPKPGPYVK	0.9665	1827.9858	Q01518
ALGWVAMAPKPGPYVK	0.9813	1827.9858	Q01518
ALGWVAMAPKPGPYVK	0.9985	1827.9858	Q01518
ALGWVAMAPKPGPYVK	0.9919	1827.9858	Q01518
ALLAHEIGFGSK	0.8130	1357.7142	Q99497
ALLPQTLLDQK	0.9875	1354.7608	P50225
ALLSAPNPDDPLANDVAEQWK	0.9995	2379.1495	P61088
ALNELLQHVK	0.9559	1279.7036	Q9Y490
ALPFFGFSEPLAAPR	0.7521	1706.8569	P22314
ALPHAILR	0.9866	977.5558	Q562R1
ALPHAILR	0.8960	977.5558	Q562R1

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
ALPHAILRLD	0.9410	1205.6669	Q562R1
ALQDMLLLK	0.9015	1175.6372	Q6P2M8-5
ALSENSGMNPIQTMTEVR	0.9337	2064.9356	P48643
ALTGGIGFIHHNCTPEFQANEVR	0.9914	2655.2401	P12268
ALVIDNGSGMCK	0.9327	1379.6247	P53505
ALVIDNGSGMCK	0.7986	1379.6251	P53505
ALVKPEVWTLK	0.9346	1426.8336	Q9UL46
ALVLLIAQEK	0.9101	1212.7230	P48637
ALVVDNGSGMCK	0.9901	1365.6169	Q562R1
ALYPEGQAPVKK	0.9532	1443.7874	P37802
ALYPEGQAPVKK	0.8664	1443.7874	P37802
ALYPEGQAPVKK	0.9874	1443.78	P37802
AMPTLIELMKDPSVVVR	0.9529	2014.0743	Q14974
AMQLLTAEIEK	0.9385	1361.7012	Q07666
ANAGPNTNGSQFFICTAK	0.9924	2012.9163	A2BFH1
ANNLVAAAIDAR	0.9037	1285.6526	P11586
ANTVLSGGTTMYPGIADR	0.9934	1910.8867	P53478
APELIHDFLVNEK	0.9998	1639.8358	P53618
APELIHDFLVNEK	0.9994	1639.8358	P53618
APMLVTGNPGVPVPAAAAAAAQK	0.9808	2217.1728	Q9NR56
APSGQPGSTK	0.7451	1044.4914	Q7VA20
APTHFLVIPK	0.9632	1237.6971	P49773
APVNVTTEVK	0.9966	1172.6115	P68103
APVNVTTEVK	0.9944	1172.6115	P68103
APVNVTTEVK	0.9779	1172.6115	P68103
AQGHGIIQVDK	0.8946	1280.6625	P29144
AQINQGESITHALK	1.0000	1624.8321	Q01518
AQINQGESITHALK	1.0000	1624.8321	Q01518
AQINQGESITHALK	0.9982	1624.8321	Q01518
AQINQGESITHALK	0.9973	1624.8247	Q01518
AQINQGESITHALK	0.9976	1624.8247	Q01518
AQLDHWALTQR	0.9909	1425.6901	Q6GTX8
ASAAAVDAGMAMAGQSPVLR	0.7298	1960.9247	P26599
ASILAAFSK	0.9047	1022.5549	P54819
ASILAAFSK	0.9742	1022.5475	P08166
ASILAAFSK	0.9910	1022.5475	P08166
ASQCQQPAENK	0.9999	1375.5938	Q01518
ASQCQQPAENK	0.9998	1375.5938	Q01518

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
ASQCQQPAENK	0.9973	1375.5938	Q01518
ASSSSLEK	0.8969	923.4348	P62736
ASSSSLEKSYELPDGQVITIGNER	0.9984	2695.3089	P62736
ASTPVFGGILSLINEHR	1.0000	1897.9798	O14773
ATAASSSSLEK	0.9265	1166.5567	P62736
ATAASSSSLEK	0.9687	1166.5493	P53478
ATAASSSSLEK	0.9407	1166.5493	P53478
ATAASSSSLEK	0.9831	1166.5493	P53478
ATAASSSSLEK	0.9845	1166.5493	P53478
ATAASSSSLEK	0.9956	1166.5493	P53478
ATAASSSSLEK	0.9602	1166.5493	P53478
ATAASSSSLEK	0.9835	1166.5493	P53478
ATAASSSSLEK	0.9931	1166.5493	P53478
ATAASSSSLEK	0.909	1166.5493	P53478
ATQLAVNKIKE	0.9939	1357.7717	Q99832
ATVLARSIAKE	0.9766	1273.7142	P10809
ATYAPVISAEK	0.9884	1264.6377	P68362
AVALAGLLAAQK	0.9982	1240.7291	P23368
AVALAGLLAAQK	0.8190	1240.7291	P23368
AVALAYGIYK	0.9922	1183.6389	O95757
AVDALIDSMSLAK	0.9998	1448.7333	P13010
AVIAELKK	0.8011	1014.6225	P10809
AVLIVAKKCPS	0.9822	1328.7638	Q04323
AVRLLLPGE	0.9342	1054.5923	Q96A08
AVTVAPPGARQGQQQAGGDGKTE	0.9997	2338.1414	Q00839-2
AVTYTEHAK	0.9752	1134.5458	P62805
AVVFGPNLLWAK	0.9989	1429.7870	Q07960
AWGLVTTAPR	0.7969	1158.5934	A6NCC3
CAEHQINLIK	0.9939	1340.6659	P25398
CAEHQINLIK	0.9312	1340.6585	Q76I81
CAGYLEGGK	0.9759	1069.4577	P00760
CAILSPAFK	0.7828	1121.5691	Q92598
CALSTSQLVACTK	0.7832	1553.7256	P54939
CAVLIVAAGVGEFEAGISK	0.9997	2006.0217	P68103
CDEGYESGFMMMKNCMDIDECQ	0.8594	2880.9926	P35555
CEDIIQLKPDVVITEK	0.9815	2043.0710	P49368
CELINALYPEGQAPVKK	0.7977	2073.0717	P37802
CELINALYPEGQAPVKK	0.9936	2073.0717	P37802
CELINALYPEGQAPVKK	0.9829	2073.0717	P37802

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
CETIIGAVP	0.8890	1046.4781	P23400
CGVDLIIGVGGGR	0.9799	1342.6377	B8GGP5
CIAIKESAK	0.9973	1162.6168	P61158
CLHFNPR	0.9989	1030.4555	P09382
CPGESSHICDFIR	0.9910	1647.6484	P45478
CPGESSHICDFIR	0.9987	1647.6484	P45478
CVLQGLQTPSCK	0.9992	1505.7044	P13489
CVVAVLPHILDTGAAGR	0.9963	1835.9464	Q15084
CVVAVLPHILDTGAAGR	0.9997	1835.9464	Q15084
DAANFEQFLQER	0.8383	1554.6851	P35268
DAFGTAHR	0.8651	961.4154	P00558
DAGAGIALNDHFVK	0.9819	1542.7579	P04406
DAGAGIALNDHFVK	0.9543	1542.7579	P04406
DAGAGIALNDHFVK	0.9958	1542.7507	P10096
DAGAGIALNDHFVK	0.9997	1542.7507	P10096
DAGAGIALNDHFVK	0.9735	1542.7507	P10096
DAGILQLVESVR	0.9993	1386.7255	P13489
DALCVLAQTVK	0.9002	1332.686	P78371
DALDKIR	0.9204	945.5032	P14625
DAMAGDFVNMVEK	0.9996	1541.6642	P10809
DANLQTLTEYLKK	0.9481	1679.8882	P55060
DANTIVCNSK	0.8726	1236.5557	P09382
DANTIVCNSKDGGAWGTEQR	0.9977	2294.0134	P09382
DANTIVCNSKDGGAWGTEQRE	0.9995	2423.0560	P09382
DAPMFVMGVNHEK	0.9848	1589.7119	P04406
DCHTAHIACK	0.9933	1327.5550	P68104
DCHTAHIACK	0.9887	1327.555	P68104
DCHTAHIACK	0.9739	1327.5550	P68104
DDHDPVDK	0.8588	1055.4308	P22626
DDVVGIVEIINSK	0.9999	1515.7933	Q01518
DEELNKLLGK	0.9704	1301.6979	P20671
DEGGFAPNILENKEGLELLK	0.9961	2329.1953	P06733
DEITYVELQKEEAQK	0.9768	1965.9683	Q00839
DEITYVELQKEEAQK	0.9986	1965.9683	Q00839
DESGPSIVHR	0.9369	1183.5370	Q9BYX7
DESGPSIVHR	0.9893	1183.537	Q9BYX7
DESGPSIVHR	0.9733	1183.5370	Q9BYX7
DESTGSIAKR	0.9650	1178.5679	P04075
DFEQEMATAASSSSLEK	0.9818	1945.8363	P60709

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
DFLLKPELLR	0.9300	1358.7710	O00148
DIAVDGEPLGR	0.9029	1228.5836	P62937
DIAVDGEPLGR	0.9126	1228.5836	P62937
DIETIGEILKK	0.8970	1401.7867	P61978
DISPQAPTHFLVIPK	0.7650	1777.9515	P49773
DISPQAPTHFLVIPK	0.9442	1777.9515	P49773
DISPQAPTHFLVIPK	0.9588	1777.9515	P49773
DIVQLPTGLTGIK	0.8570	1469.8242	P34932
DKANAQAAALYK	0.9998	1406.7306	P40121
DKANAQAAALYK	0.9801	1406.7306	P40121
DKANAQAAALYK	0.9953	1406.7306	P40121
DKDGDGTITTK	0.9540	1293.6201	P62158
DKDGDGTITTKE	0.9500	1422.6627	P62158
DKFDENAK	0.9956	1109.5141	P00558
DKGLQTSQDAR	0.9488	1333.6374	P27797
DKLNVITVGPR	0.9931	1326.7408	P04040
DKLNVITVGPR	0.9634	1326.7408	P04040
DKPLKDVIIAD	0.8905	1369.7605	P23284
DKYLIPNATQPESK	0.9966	1746.894	P31946
DKYLIPNATQPESK	0.8435	1746.8940	P31946
DLCHALR	0.9924	971.4395	P24534
DLFNAVGDGIVLCK	0.9893	1635.8079	P13796
DLFNAVGDGIVLCK	0.9393	1635.8079	P13796
DLVVGLCTGQIK	0.9785	1417.7388	P06733
DMVPGKPMCVESFSDYPPLGR	0.9984	2497.1228	P68104
DMVPGKPMCVESFSDYPPLGR	0.9872	2497.1228	P68104
DMVPGKPMCVESFSDYPPLGR	0.9988	2497.1228	P68104
DNDIMLIK	0.9709	1076.5324	P35030
DNSSRPSQVVAETR	0.8388	1632.7604	P13639
DNVICPGAPDFLAHVR	0.9995	1867.8788	P21964
DQANLTVK	0.7675	1003.5087	P09382
DQAQKAEGAGDAK	0.8455	1431.6742	P05204
DQIQNAQYLLQNSVK	0.9949	1876.9431	P61978
DQLHAAVGASR	0.9783	1211.5795	P13804
DQSYKPDENEVR	0.9939	1594.7011	P31939
DRTVIDYNGER	0.9844	1424.6432	P07237
DRTVIDYNGERTLD	0.9619	1753.8019	P07237
DSCTCAGSCKCKE	0.9934	1705.6317	P02795
DSLLAGPVAEYLK	0.9980	1490.7769	Q01518

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
DSLLAGPVAEYLK	0.7829	1490.7769	Q01518
DSLLAGPVAEYLK	0.9357	1490.7769	Q01518
DSLLAGPVAEYLK	0.7614	1490.7697	Q01518
DSLLAGPVAEYLK	0.8376	1490.7697	Q01518
DSLLAGPVAEYLK	0.9653	1490.7695	Q01518
DSLYVEKIDVGEAEPR	0.8656	1934.9373	P54577
DSYVGDEAQSKR	0.9955	1469.6535	P62736
DTFLEHMCR	0.9732	1295.5175	P14618
DTFLEHMCR	0.7707	1295.5175	P14618
DTFWKEFGTNIK	0.8257	1628.7987	Q58FF3
DTKPGTTGSGAGSGGPGGLTSAAPAGGDKK	0.9999	2728.3417	P67809
DTKPGTTGSGAGSGGPGGLTSAAPAGGDKK	0.8637	2728.3417	P67809
DTKPGTTGSGAGSGGPGGLTSAAPAGGDKK	0.9915	2728.3417	P67809
DTLLVDVEPK	0.9029	1243.6448	P62314
DTYNCDLHFK	0.9823	1427.5928	Q9BUJ2
DTYNCDLHFK	0.8452	1427.5928	Q9BUJ2
DTYNCDLHFK	0.9124	1427.5928	Q9BUJ2
DVCPLTLGIETVGGVMTK	0.9905	2004.9937	Q91883
DVVVLPGGNLGAQNLSESAAVK	1.0000	2253.1753	Q99497
DVVVLPGGNLGAQNLSESAAVK	0.9864	2253.1753	Q99497
DVVVLPGGNLGAQNLSESAAVK	0.9439	2253.1753	Q99497
DVVYALK	0.9718	922.4912	P62805
DVVYALKR	0.8879	1078.5923	P62805
DYNGHVGLGVK	0.9845	1273.6203	P15880
DYNGHVGLGVK	0.9561	1273.6203	P15880
EALAAAELLKK	0.9997	1299.7550	P29401
EAPNPKL	0.8195	883.4477	Q39072
EAPNPKL	0.8939	883.4477	Q39072
EASEAYLVGLFEDTNLCAIHAK	0.7360	2566.2162	Q71DI3
EASGGGAFLVLPLGK	0.9999	1530.8117	P31146
EAYLVGLFEDTNLCAIHAK	0.9475	2279.1044	P68431
EAYLVGLFEDTNLCAIHAK	1.0000	2279.1044	P68431
EDTNLCAIHAK	0.9986	1386.6350	P68431
EDTNLCAIHAK	0.9858	1386.635	P68431
EDTNLCAIHAK	0.9785	1386.6350	P68431
EDTNLCAIHAK	0.9989	1386.6276	P84227
EDTNLCAIHAK	0.9998	1386.6276	P84227
EDTNLCAIHAK	0.9998	1386.6277	P84227
EDTNLCAIHAK	0.9990	1386.6277	P84227

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
EGNDYFKEK	0.7990	1272.5775	O95801
EHGIQPDGQMPSDK	0.9983	1653.7205	Q71U36
EIAPHALLQAVLK	0.9996	1517.8718	P49327
ELDQDMVTEDEDDPG	0.7924	1792.6297	Q9NRL2
ESCGIHETTFNSIMK	0.9994	1868.8185	P60709
ESCGIHETTFNSIMK	0.9991	1868.8185	P60709
ESHIQSTSDR	0.7984	1246.5326	Q14974
EVAAAVPAPK	0.9357	1049.5584	Q82KE2
EVGDIMLIR	0.9609	1132.5699	P00491
EVGVLVGK	0.9975	915.5178	P07737
EVLLPGLQK	0.8901	1111.6389	P07954
FAALTSIAQK	0.8735	1164.6291	Q9Y5Y2
FAEALAAHK	0.8814	1072.5453	P07237
FAEALAAHK	0.9990	1072.5453	P07237
FAEALAAHK	0.9996	1072.5453	P07237
FAGILSQGLR	0.9941	1148.6090	P09874
FAGPHAALANK	0.9990	1211.6199	Q9BY44
FAGSVPPP	0.7937	858.395	P13002
FALLEIPK	0.8483	1045.5960	O94915
FAPVNVTTEVK	0.9611	1319.6874	P68104
FAPVNVTTEVK	0.9868	1319.6799	P68103
FAPVNVTTEVK	0.9994	1319.6797	P68103
FAPVNVTTEVK	0.8600	1319.6797	P68103
FAPVNVTTEVK	0.9995	1319.6797	P68103
FAQINQGESITHALK	0.9989	1771.9005	Q01518
FAQINQGESITHALK	0.9986	1771.9005	Q01518
FAQINQGESITHALK	0.9997	1771.9005	Q01518
FAQINQGESITHALK	0.9962	1771.9005	Q01518
FAQINQGESITHALK	0.9999	1771.8927	Q01518
FCAILHR	0.8255	1003.4810	Q8N3D4
FCSEYRPK	0.9160	1201.5338	P09429
FDQANLTVK	0.8275	1150.5696	P09382
FDSLLAGPVAEYLK	0.9904	1637.8377	Q01518
FDSLLAGPVAEYLK	0.9966	1637.8377	Q01518
FEDTNLCAIHAK	0.9996	1533.7034	P68431
FEDTNLCAIHAK	0.9999	1533.7034	P68431
FEDTNLCAIHAK	0.9998	1533.7034	P68431
FEDTNLCAIHAK	0.9621	1533.6957	P84227
FEDTNLCAIHAK	0.9551	1533.6957	P84227

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
FEQEMATAASSSSLEK	0.9996	1830.802	P53478
FEQEMATAASSSSLEK	0.9994	1830.802	P53478
FFVQTCR	0.9714	1044.4599	B2RPK0
FFVQTCREE	0.9390	1302.5451	B2RPK0
FGGGVIGDLAGFAAANYLR	0.7336	1955.9567	Q1LU62
FGILLDQGQLNK	0.9081	1460.7776	Q00610
FGPDICGPGTK	0.9951	1263.5632	Q4VIT5
FGTHETAFLGPK	0.9962	1419.6935	P51858
FGVLGLDLWQVK	0.9996	1489.8081	P27797
FGVLGLDLWQVK	0.9581	1489.8007	Q4VIT5
FGYFEVTHDITK	0.9994	1571.7409	P04040
FGYFEVTHDITK	0.9974	1571.7337	Q2I6W4
FHTEQMYK	0.8022	1198.5229	P59998
FIAIKPDGVQR	0.8202	1358.7459	P15531
FIFIDSDHTDNQR	0.9355	1694.7437	P07237
FIGAIAIGDLVK	0.9927	1331.7601	P78371
FIGAIAIGDLVK	0.9997	1331.7527	Q3ZBH0
FIGNSTAIQELFK	0.9926	1582.8143	Q13885
FIGNSTAIQELFK	0.9787	1582.8067	Q9YHC3
FIGNSTAIQELFKR	0.8744	1738.9154	Q13885
FIGNSTAIQELFKR	0.9711	1738.9154	Q13885
FILFKDAASVEK	0.9951	1510.8184	Q99729
FIVLTTSAGIMDHEEAR	0.9959	1976.9337	Q9LX88
FLAAGLK	0.8607	834.4752	Q9ULV1
FLAQLKDECPEVR	0.9664	1719.8402	P30153
FLGMESCGIHETTFNSIMK	1.0000	2317.0329	P60709
FLGMESCGIHETTFNSIMK	0.9998	2317.0329	P60709
FLLNTLQENVNK	0.9991	1547.8096	Q9HAV4
FLLPHPGLQVATSPDFDGK	0.9734	2154.0898	Q6DD88
FLLPHPGLQVATSPDFDGK	0.9852	2154.0898	Q6DD88
FLNLANDPTIER	0.9938	1489.7313	P15313
FLPEFLVSTQK	0.9975	1423.7500	P30740
FLPEFLVSTQK	0.7975	1423.75	P30740
FLPVIGLVDAEK	0.9976	1415.7812	P17980
FLPVIGLVDAEK	0.9904	1415.7812	P17980
FLTTGVLSTLR	0.7258	1294.7033	P49915
FMNTELAAFTK	0.8877	1387.6594	P31949
FMPGFAPLTSR	0.9956	1310.6156	Q9YHC3
FMVVNDAGRPK	0.9649	1348.671	P11142

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
FMVVNDAGRPK	0.9975	1348.6710	P11142
FMVVNDAGRPK	0.9924	1348.6710	P11142
FNTLQTK	0.7459	966.4923	P12814
FNVINGGSHAGNK	0.9996	1429.6851	P06733
FPRPVTVEPMDQLDDEEGLPEK	0.7247	2656.2479	Q15233
FQLAPAILQGQTK	0.7493	1529.8354	Q96JB5
FSAPKPQTSPSPK	0.9924	1514.7881	Q01518
FSGLFGGSSK	0.9399	1101.5243	P54920
FSTPLLLGKK	0.9838	1246.7437	P40926
FSTPLLLGKK	0.8605	1246.7437	P40926
FTTTAERE	0.9329	1041.4515	Q562R1
FTVWDVGGQDK	0.8143	1366.6306	P84077
FVALSTNTTKVKE	0.8709	1580.8562	P06744
FVLDEFKR	0.7413	1168.6029	P26641
FVMGVNHEK	0.9932	1175.5546	P04406
FVTFCTK	0.9301	1017.4742	O60506
FVTFDDHDPVDK	0.9906	1549.6838	P22626
FVTFDDHDPVDK	0.9648	1549.6838	P22626
FVVEVIK	0.8488	948.5433	Q07021
FYELSENDLNFIK	0.9989	1746.8253	P13639
FYFDPLINPISHR	0.8857	1705.8365	Q6P2Q9
FYVNGLTLGGQK	0.8321	1411.7248	P07737
FYVNGLTLGGQK	0.9129	1411.7248	P07737
FYVNGLTLGGQK	0.9978	1411.7248	P07737
FYVNGLTLGGQK	0.9839	1411.7248	P07737
GAAAAIEAAAK	0.9935	1058.5508	Q9Y490
GAAGVMAIEHADFAGVER	0.9987	1887.8686	P78371
GAAVQAAILSGDK	0.9963	1315.6884	P11142
GAAVSAGHGLPAK	0.9992	1250.6520	O75367
GAFLHIK	0.9826	900.4970	Q8WUM4
GAGLMGAGIAQVSVDK	0.9985	1588.7957	P40939
GAGLMGAGIAQVSVDKGLK	0.9483	1915.0349	P40939
GAGNPVGDKLNVITVGPR	0.9997	1879.0064	P04040
GALAIANFAR	0.9940	1090.5671	P52306
GALLEEAEQLLDR	0.9389	1543.7556	P48643
GALLPCEECSGQLVFK	0.9964	1922.8944	P18493
GAPFLKEGASEEEIR	0.8969	1747.8529	P23141
GASGGAYEHR	0.9997	1091.4532	P31943
GASGGAYEHR	0.9912	1091.4532	P31943

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
GASIYIENKEEK	0.9465	1523.7619	O75832
GASTGIYEALELR	0.9981	1466.7079	Q9C9C4
GDDDSGPGPK	0.8628	1059.4257	Q13316
GDLDETSSNEGGVK	0.7461	1522.6457	Q9SLF3
GDRFTDEEVDELYR	0.9128	1830.7809	P19105
GDRFTDEEVDELYR	0.9892	1830.7734	P19105
GDRFTDEEVDELYR	0.9886	1830.7737	P19105
GEKPVGSLAGIGEVLGK	0.9924	1753.9726	O75531
GESITHALK	0.9755	1070.5508	Q01518
GESITHALK	0.9604	1070.5508	Q01518
GFPCNQFGHQENAK	0.9998	1748.7477	P07203
GFSAFPFELLHTPEK	0.9999	1834.9042	P07099
GFSAFPFELLHTPEK	0.9911	1834.9042	P07099
GGPLPPHLALK	0.8714	1214.6924	Q08211
GGQDILSMMGQLMKPK	0.9999	1876.9361	Q9Y265
GGQDILSMMGQLMKPK	0.9995	1876.9361	Q9Y265
GGSHAGNKLAMQE	0.9427	1414.6411	P06733
GGSSEPCALCSLHSIGK	0.8927	1874.8403	P14174
GGSSEPCALCSLHSIGK	0.8867	1874.8403	P14174
GGTTMYPGIADR	0.9183	1325.5822	P62736
GGTTMYPGIADRMQKE	0.9909	1869.8501	P62736
GGTTMYPGIGER	0.8812	1325.5748	P26183
GGTTMYPGIGER	0.9474	1325.5748	P26183
GGVCEPLK	0.8395	974.4644	Q9UHF7
GGVLPNIQAVLLPK	1.0000	1533.9031	Q96QV6
GGVLPNIQAVLLPK	0.9999	1533.9031	Q96QV6
GGVLPNIQAVLLPK	0.9530	1533.8957	P04908
GGVLPNIQAVLLPK	0.9907	1533.8957	P04908
GGVLPNIQAVLLPK	0.9987	1533.8957	P04908
GGVLPNIQAVLLPK	0.9987	1533.8957	P04908
GGVLPNIQAVLLPK	0.9920	1533.8957	P04908
GGVMSGAVPAAAAQEAVEEDIPIAK	0.7462	2496.2318	P52815
GHILAAEQLSR	0.8893	1281.6577	Q15738
GHLLLVAK	0.9794	965.5810	Q9BX68
GHPAFVNYSTSQK	0.9998	1550.7266	P14866
GHPLGASGCR	0.9845	1098.4777	Q9BWD1
GHSVEELCK	0.9973	1173.5236	P29401
GHVGADLAALCSEAALQAIR	0.9999	2110.0377	P55072
GHVLAAGCGQNPVR	0.9999	1522.7211	Q9BWD1

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
GHVLAAGCGQNPVR	0.9996	1522.7211	Q9BWD1
GHVLAAGCGQNPVR	0.9924	1522.7211	Q9BWD1
GIFVEKYDPTIEDSYR	0.9973	2046.9687	P62834
GIHETTFNSIMK	0.9991	1492.7057	P53478
GIHETTFNSIMK	0.9998	1492.7057	P53478
GIHETTFNSIMK	0.9997	1492.7057	P53478
GIHETTFNSIMK	0.9990	1492.7057	P53478
GIHETTFNSIMK	0.9998	1492.7058	P53478
GIHETTFNSIMK	0.9523	1492.7132	Q562R1
GIHETTFNSIMK	0.9967	1492.7132	Q562R1
GIHETTFNSIMK	0.9343	1492.7132	Q562R1
GIMNSFVNDIFER	0.9978	1628.7327	P06900
GIMNSFVNDIFER	0.9998	1628.7327	P06900
GIPVLVLGNKR	0.9742	1280.7717	Q96BM9
GIPVLVLGNKR	0.9663	1280.7717	Q96BM9
GIPYLDAPSEAEASCAALVK	0.9998	2177.0462	P39748
GISLANLLLSK	0.9304	1243.7288	P08397
GISQGLADNTVIAK	0.9995	1501.7888	P26639
GITLPVDFQGR	0.9878	1289.6442	P31943
GITLPVDPEGK	0.8388	1240.6377	Q5E9J1
GIVPIVEPEILPDGDHDLK	0.9999	2171.1262	P04075
GLAWSKTGPVAKE	0.9915	1486.7932	Q01518
GLDLWQVK	0.9439	1073.5658	P27797
GLEVFHAGTALK	0.9976	1357.7142	P22102
GLFPCVDELSDIHTR	0.9869	1845.8468	Q92974
GLIFVVDSNDR	0.8725	1321.6415	P84077
GLLWALEPEKPLVR	0.9991	1735.9773	Q86TX2
GLLWALEPEKPLVR	0.9986	1735.9773	Q86TX2
GLTHTAVVPLDLVK	0.8956	1577.8929	Q00325-2
GLTLGGQKCSVIRD	0.9984	1618.8249	P07737
GLVASNLNLKPGECLR	0.9770	1855.9726	P09382
GMGMEGIGFGINK	0.9787	1425.6533	P52272
GMILPTMNGESVDPVGQPALK	0.9517	2269.1235	O95433
GNIVGLVGVDQFLVK	0.8504	1672.9227	Q3SYU2
GNIVGLVGVDQFLVK	0.9997	1672.9227	Q3SYU2
GNIVGLVGVDQFLVK	0.9982	1672.9227	Q3SYU2
GPKPALPAGTEDTAK	1.0000	1595.8307	P06396
GPKPALPAGTEDTAKEDAANR	0.8412	2252.1185	P06396
GQDEMIDVIGVTK	0.8514	1519.7341	P39023

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
GQKDSYVGDEAQSK	0.9999	1654.7587	P62736
GQLLTSSNYDDDEKK	0.9999	1855.8588	P11388
GQLLTSSNYDDDEKK	0.9998	1855.8588	P11388
GQLLTSSNYDDDEKK	0.9476	1855.8588	P11388
GQLMNMLSHPVIR	0.9825	1582.7860	Q7Z6Z7
GQSGAGNNWAK	0.9963	1204.5373	Q13885
GQSGAGNNWAK	0.9661	1204.5373	Q13885
GQSGAGNNWAK	0.8827	1204.5299	Q9YHC3
GQVITIGNER	0.9861	1173.5890	P62736
GSLGQGLGAACGMAYTGK	0.9989	1813.8239	P29401
GSPKADSPGSLTI	0.7936	1344.6673	Q8NFW5
GSTSDLGHCEK	0.9672	1305.5408	P22234
GSTSDLGHCEK	0.9868	1305.5408	P22234
GTAAVALAGLLAAQK	0.9999	1469.8354	P23368
GTFALNLLK	0.9662	1091.6127	P35237
GTQDQIQNAQYLLQNSVK	0.9924	2163.0708	P61978
GVDLLADAVAVTMGPK	0.9054	1671.8654	P10809
GVDLLADAVAVTMGPK	0.9908	1671.8577	P10809
GVGYLAGCLVHALGEK	0.9997	1758.8875	Q9Y3Z3
GVGYLAGCLVHALGEK	0.9992	1758.8875	Q9Y3Z3
GVHQVPTENVQVHFTER	0.9582	2063.9926	Q9HB71
GVLPNIQAVLLPK	0.8899	1476.8816	Q96QV6
GVPMPDKYSLEPVAVELK	0.9709	2115.1074	P00558
GVPMPDKYSLEPVAVELK	0.8261	2115.1074	P00558
GVSHPVLK	0.8490	951.5290	P22695
GVSLAVCK	0.9987	948.4851	P06733
GVSLAVCK	0.8683	948.4851	P06733
GVSLQELNPEMGTDNDSENWK	1.0000	2478.0757	P07195
GVVVLHK	0.9855	866.5126	O43395
GWVAMAPKPGPYVK	0.8966	1643.8646	Q01518
HACIGGTNVR	0.7667	1171.5305	P60842
HAQVADMK	0.9908	1014.4705	P35579
HCASPPPSSNNK	0.9983	1410.6098	Q15738
HCIDPNDSK	0.9925	1200.4982	Q15185
HEALAAAELLKK	0.9990	1436.8139	P29401
HEALAAAELLKK	0.9717	1436.8139	P29401
HEKYDNSLK	0.8571	1276.6200	P04406
HHTFYNELR	0.8326	1303.5846	Q9BYX7
HIISSNLEK	0.9587	1155.6036	Q9UL46

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
HILSPWGAEVK	0.9009	1351.7037	P09874
HLATGDMLR	0.9014	1100.5185	P54819
HLATGDMLR	0.9927	1100.5185	P54819
HLKSPVR	0.8036	951.5402	Q9UIF9
HLLLQNNLPAVR	0.9991	1474.8156	P48643
HLQLAIRNDEE	0.9998	1424.6796	Q96QV6
HNCAVEFNFGQK	0.9986	1565.6833	Q00839
HQALLGTIR	0.9985	1095.5937	P25705
HQATILPK	0.9370	1022.5661	P49327
HQATILPK	0.9952	1022.5661	P49327
HQGVMVGMGQKDSYVGDE	0.9993	2051.8829	P62736
HSFGGGTGSGFTSLLMER	1.0000	1927.8635	Q71U36
HSGIAPR	0.7955	824.4041	Q8IWT3
HSIIETLR	0.9792	1055.5512	P07900
HSIVLPLK	0.8109	1021.6072	O94979
HSLGGGTGSGMGTLLISK	0.8678	1787.8988	Q13885
HSLGGGTGSGMGTLLISK	0.9789	1787.8988	Q13885
HSLLPALCDSK	0.7387	1355.6581	A1YES6
HTISPLDLAK	0.9779	1209.6506	Q15365
HTVLPEALER	0.9130	1251.6286	Q3B7M9
HTVPIYEGYALPHAILR	0.7847	2037.0584	P60709
HVTYAGAAVDELGK	0.9999	1545.7576	P30086
HVVNIGAEDLK	0.9985	1309.6778	P13796
IAAQYSGAQVR	0.9993	1250.6156	P26641
IADLVVGLCTGQIK	0.9561	1601.8599	P06733
IAGHPAFVNYSTSQK	0.9148	1734.8478	P14866
IAIGDLVK	0.8966	943.5491	P78371
IANLFNR	0.9812	934.4773	P13796
IAPALVSKKLNVTE	0.8933	1625.9504	P06733
IAPALVSKKLNVTE	0.9997	1625.9504	P06733
IAPIVIFASNR	0.9998	1287.7087	Q9Y265
IAQGGVLPNIQAVLLPK	1.0000	1846.0828	Q96QV6
IAQGGVLPNIQAVLLPK	0.9991	1846.0757	P04908
IAQGGVLPNIQAVLLPK	0.9981	1846.0757	P04908
IASGGVLPNIHPELLAK	0.9989	1844.0308	O75367
IASGGVLPNIHPELLAK	0.9897	1844.0308	O75367
ICAGPTALLAHEIGFGSK	0.7868	1956.988	Q99497
ICQQNGIVPIVEPEILPDGDHDLK	0.9405	2814.4010	P04075
ICQQNGIVPIVEPEILPDGDHDLK	0.9995	2814.401	P04075

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
IFIDSDHTDNQR	0.9968	1547.6753	P07237
IFIDSDHTDNQR	0.9920	1547.6753	P07237
IFIDSDHTDNQR	0.9775	1547.6679	P05307
IGAIAIGDLVK	0.9970	1184.6917	P78371
IGNSTAIQELFKR	0.9922	1591.8470	Q13885
IGRRFDDAVVQSD	0.9438	1564.7382	P11142
IGSLICNVGAGGPAPAAGAAPAGGPAPSTAAAPAEEK	0.9994	3341.6462	P05386
IGYPITLFVEK	0.9997	1394.7598	Q14568
IGYPITLYLEK	0.9793	1424.7703	Q58FF7
IIAEGIPEALTR	0.8924	1369.7353	P53396
IINSLYKNKE	0.8141	1364.7452	P14625
IIRPRPPK	0.7241	1091.6716	Q16881-2
IKKIGYNPD	0.8868	1190.6447	P68104
ILGQNGISDLVK	0.9844	1371.751	P00338
ILGQNGISDLVK	0.8321	1371.7437	P00338
ILGQNGISDLVK	0.9640	1371.7436	P00338
ILGQNGISDLVK	0.7707	1371.7437	P00338
ILGTTLKDEGK	0.9996	1317.7292	O75083
ILGTTLKDEGK	0.8954	1317.7292	O75083
ILLVQPTKRPE	0.9959	1408.8190	P84090
ILNVSAVDKSTGKE	0.9953	1603.8569	P11142
ILTHGIFSGPAISR	0.9878	1555.8259	P60891
ILTLKYPIE	0.9520	1204.6855	P62736
IMNSFVNDIFER	0.9996	1571.7117	P06900
IMNSFVNDIFER	0.9972	1571.7117	P06900
IMNSFVNDIFER	0.9804	1571.7117	P06900
IMNSFVNDIFER	0.8862	1571.7117	P06900
INGHNAEVR	0.9906	1096.5162	P22626
IPNEIIHALQAGR	0.9915	1518.8055	P52272
IPTEGGDFNEFPVPEQFK	0.9989	2166.0058	Q01518
IPTLITQLTQK	0.9882	1370.7921	P55060
IQANPLLEAFGNAK	0.9970	1600.8361	Q9UKX3
IQAVLLPK	0.7468	996.6120	Q96QV6
IQAVLLPK	0.8108	996.6120	Q96QV6
IQAVLLPK	0.9811	996.6120	Q96QV6
IQAVLLPK	0.9327	996.612	Q96QV6
IQAVLLPKKTE	0.9975	1382.8285	Q96QV6
IQGITKPAIR	0.9891	1211.7138	P62805
IQWITTQCR	0.9828	1292.6084	P37802

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
IRNDEELNK	0.9718	1245.6101	Q96QV6
ISADIETIGEILKK	0.9999	1672.9399	P61978
ISADIETIGEILKK	0.9899	1672.9399	P61978
ISADIETIGEILKK	0.8749	1672.9399	P61978
ISATLPHEILEMTNK	0.9952	1811.9239	P38919
ISKLIFKS	0.8192	1078.6539	Q9ULI3
ISLCQAILDETKGDYEK	0.9997	2126.0354	P04083
ISLCQAILDETKGDYEK	0.9999	2126.0354	P04083
ISPYFINTSKGQKCE	0.9980	1914.9298	P10809
ISRMQYAPNTQVE	0.9635	1623.7463	P40121
ITALHIK	0.9859	910.5388	P62263
ITFDQANLTVK	0.8732	1364.7088	P09382
ITFDQANLTVK	0.9371	1364.7017	P09382
ITGKTFSSR	0.8278	1111.5774	O15144
ITWIGENVSGLQR	0.9387	1559.7844	Q14019
ITYTDEEPVKK	0.9329	1465.7453	O14979
IVASKASLRE	0.9035	1188.6614	P13489
IVCNSKDGGAWGTEQRE	0.9993	2021.9013	P09382
IVGNSALK	0.7209	916.5130	Q9Y283
IVPALEIANAHR	0.9979	1390.7469	P10809
IVPALEIANAHR	0.9220	1390.7469	P10809
IVPALEIANAHR	0.9801	1390.7397	P10809
IVPALEIANAHR	0.9661	1390.7395	P10809
IVPALEIANAHR	0.9979	1390.7397	P10809
IVPIVEPEILPDGDHDLK	0.9998	2114.1048	P04075
IVPIVEPEILPDGDHDLK	0.9759	2114.1048	P04075
IVPTGKTGLIIGKGGE	0.9981	1682.9719	Q96AE4
IVSWGSGCAQK	0.7351	1307.6006	P00760
IVSWGSGCAQK	0.9912	1307.6006	P00760
IVVIGHVDSGK	0.9998	1238.6771	P68104
IYTNYEAGKDDYVK	0.9554	1821.8573	P09211
KADGIVSKNF	0.8101	1221.6506	P63220
KANAQAAALYK	0.9593	1291.7036	P40121
KAPLDIPVPDPVKE	0.8980	1660.9188	Q06323
KDDAMLLK	0.9173	1076.5688	P10809
KDSPSVWAAVPGK	0.9997	1484.7776	P07737
KDSTLIMQLLR	0.9913	1432.7860	P31946
KEPAVLELEGK	0.9973	1355.7374	Q01518
KEPISVSSEQVLK	0.9982	1586.8668	P00918

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
KGTVAVQEK	0.8271	1102.6135	Q8NHV4
KLAPGELTIIL	0.7571	1282.7648	P04114
KNNQITNNQR	0.9946	1344.6646	P00558
KPGMVVTFAPVNVTTEVK	0.9903	2060.1129	P68104
KPIIDLYEEMGK	0.9718	1578.8115	P30085
KPISVEGSSK	0.8588	1174.6346	Q9HB71
KPMCVESFSDYPPLGR	0.9997	1997.9127	P68104
KQDLPNAMNAAEITDK	0.9861	1901.9305	P84077
KQGQDNLSSVKE	0.9982	1475.7368	P30040
KSESEILR	0.7890	1076.5614	Q05823
KVEFLECSAK	0.9876	1353.6751	Q9Y5M8
LAAAELLKK	0.9956	1099.6753	P29401
LAALGGNSSPSAKD	0.8290	1402.6840	P05387
LAALGGNSSPSAKD	0.9960	1402.6840	P05387
LAASALPALVMSK	0.9913	1386.7693	P36578
LAAVGLVGDLCR	0.9969	1330.6737	Q14974
LACNIALDAVK	0.8981	1302.6754	P49368
LADNVICPGAPDFLAHVR	0.9244	2051.9999	P21964
LAGGIIGVK	0.9751	942.5650	P61978
LAGLATDVQTVAQR	0.9988	1529.7950	P49720
LAGPVAEYLK	0.9981	1175.6338	Q01518
LAGPVAEYLK	0.8948	1175.6338	Q01518
LAGPVAEYLK	0.9987	1175.6338	Q01518
LAGPVAEYLK	0.9994	1175.6338	Q01518
LAGPVAEYLK	0.9601	1175.6264	Q01518
LAGPVAEYLK	0.9645	1175.6264	Q01518
LAHILSPWGAEVK	0.9888	1535.8248	P09874
LAHILSPWGAEVK	0.9823	1535.8248	P09874
LAIIDPGDSDIIR	0.9914	1484.7623	P62888
LAIIDPGDSDIIR	0.8758	1484.7623	P62888
LAIVEALNGK	0.8556	1142.6447	Q99497
LAIVEALNGKEVAAQVK	0.9999	1896.0832	Q99497
LANHSLR	0.9446	897.4569	P13489
LANSLACQGKYTPSGQAGAAASE	0.9993	2367.0913	P04075
LAPLPPLPAQFK	0.8083	1406.8074	Q9NP79
LAPVNIFK	0.7956	1016.5807	P78371
LAQYLINAR	0.9931	1148.6090	Q15365
LASYAVQSK	0.9245	1081.5556	P26038
LATYAPVISAEK	0.9423	1377.7217	P68362

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
LATYAPVISAEK	0.9646	1377.7217	P68362
LATYAPVISAEK	0.9575	1377.7217	P68362
LAVAVGHVK	0.9947	1008.5869	P62906
LAVDAVIAELK	0.9999	1256.7128	P10809
LAVDAVIAELK	1.0000	1256.7128	P10809
LAVDAVIAELK	0.9984	1256.7128	P10809
LAVDAVIAELK	0.9916	1256.7054	P10809
LAVDAVIAELKK	0.9999	1412.8391	P10809
LAVDAVIAELKK	0.9998	1412.8391	P10809
LAVDAVIAELKK	0.9994	1412.8391	P10809
LAVDAVIAELKK	0.8569	1412.8317	P10809
LAVDAVIAELKK	0.9021	1412.8317	P10809
LAVDAVIAELKK	0.9383	1412.8317	P10809
LAVLQQFK	0.9874	1061.6022	O60506
LCAIHAK	0.9969	927.4748	P68431
LCAIHAK	0.8768	927.4748	P68431
LCKPEPELNAAIPSANPAK	0.9960	2163.1146	Q8WUM4
LEGGKQPR	0.8731	999.5250	Q9UKZ4
LGALTLPLAR	0.9873	1111.6501	Q9BSJ8
LGALTPMPAVR	0.7633	1228.6386	Q9UHC9
LGGPEAAKSDETAAK	0.9994	1587.7892	P04792
LGIPFAKPPLGPLR	0.8858	1590.9398	P23141
LGSLALYEK	0.9909	1108.5916	P36542
LGSLALYEK	0.9095	1108.5842	P36542
LIANGPTGPVSF	0.9292	1259.6224	C7G046
LINIIPEDHIPLNLSGK	0.9754	2001.1047	O95602
LIPHDFGMK	0.7430	1172.5800	P09874
LIQTADQLR	0.9695	1144.5988	P18031
LISAGLPPLK	0.8619	1123.6753	O43143
LISFGAAGPPR	0.9936	1172.6090	Q75VX8
LISVYSEKGESSGK	0.9401	1626.8253	P36578
LITNFHTEQMYK	0.9897	1639.7816	P59998
LIVLEGVDR	0.9916	1100.5978	P23919
LIVLEGVDR	0.9763	1100.5978	P23919
LIYTNYEAGKDDYVK	0.9982	1934.9414	P09211
LIYTNYEAGKDDYVK	0.9990	1934.9414	P09211
LIYTNYEAGKDDYVK	0.9993	1934.9414	P09211
LIYTNYEAGKDDYVK	0.9975	1934.9337	P09211
LIYTNYEAGKDDYVK	0.9992	1934.9337	P09211

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
LKEDQTEYLEER	0.9785	1667.7790	Q58FF7
LKQGQDNLSSVKE	0.9061	1588.8209	P30040
LLAAEFLK	0.9988	1019.5803	Q99832
LLAAEFLK	0.9987	1019.5803	Q99832
LLAALGGNSSPSAKD	0.9988	1515.7681	P05387
LLAALGGNSSPSAKD	0.9889	1515.7681	P05387
LLAGIECPR	0.9662	1115.5545	Q7KZF4
LLAGPVAEYLK	0.9447	1288.7179	Q01518
LLAGPVAEYLK	0.9722	1288.7179	Q01518
LLAGPVAEYLK	0.8798	1288.7105	Q01518
LLAGPVAEYLK	0.9081	1288.7105	Q01518
LLAGPVAEYLK	0.8815	1288.7105	Q01518
LLALEPELEAR	0.8590	1340.7088	P04843
LLAYTLGVK	0.8631	1092.6331	P68104
LLLAGVFR	0.8367	975.5654	Q9Y678
LLLLVGGVDQSPR	1.0000	1453.8041	P33993
LLSLAAAAK	0.7301	972.5756	P0CAP2-3
LLSQNLVVKPDQLIK	0.9543	1851.0982	P53396
LNGIPGLER	0.8637	1055.5512	Q8NH56
LNSQKAGKE	0.8945	1117.5880	Q13185
LNVVDIAGLVK	0.9985	1255.7288	Q9NTK5
LQEYVANLLK	0.9991	1305.7081	O14980
LQGIPVLVLGNK	0.8000	1365.8132	Q96BM9
LQGIPVLVLGNKR	0.9928	1521.9143	Q96BM9
LQGIPVLVLGNKR	0.9952	1521.9143	Q96BM9
LQGIPVLVLGNKR	0.9836	1521.9143	Q96BM9
LQGVDLLADAVAVTMGPK	0.9998	1913.0080	P10809
LQGVDLLADAVAVTMGPK	0.9998	1913.0006	P10809
LQLAIRNDE	0.9724	1158.5781	Q96QV6
LQLAIRNDEE	0.9876	1287.6207	Q96QV6
LQPLLDNQVGFK	0.9129	1486.7932	P20618
LQTVAKNKDQGTYE	0.9444	1737.8686	P60660
LSAGGAAVGGRR	0.7299	1158.6006	Q6ZSJ9
LSASFEPFSNK	0.9976	1341.6353	P27797
LSFMNTELAAFTK	0.9941	1587.7677	P31949
LSLAEAQLR	0.9913	1087.5774	Q86UX7
LSPLAAAVGGVASQEVLK	0.9997	1825.0097	A0AVT1
LSVPCILGQNGISDLVK	0.9992	1928.0189	P00338
LVADENPFAQGALK	0.9999	1587.8045	P06396

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
LVANVTNPNSTEHMK	0.8630	1769.8519	Q14974
LVASNLNLKPGECLR	0.9993	1798.9512	P09382
LVFDDVVGIVEIINSK	0.9997	1875.0067	Q01518
LVFLPFADDKR	0.9573	1435.7612	P12956
LVFLPFADDKR	0.9727	1435.7612	P12956
LVFLPFADDKR	0.9609	1435.7612	P12956
LVGAGAIGCELLK	0.9757	1415.7595	P22314
LVGAGAIGCELLK	0.9853	1415.752	P31254
LVGLFEDTNLCAIHAK	0.9965	1915.9614	P68431
LVGLFEDTNLCAIHAK	0.9983	1915.9614	P68431
LVGLIQK	0.9638	885.5436	P17655
LVHWNTK	0.8963	1012.5242	P00918
LVLTDPDAPSRK	0.7857	1426.7568	P30086
LVSSSADPEGHFETPIWIER	0.9745	2357.1076	Q14697
LVTASQCQQPAENK	0.9306	1688.7940	Q01518
LVTASQCQQPAENK	0.9619	1688.7940	Q01518
LVTASQCQQPAENK	0.9994	1688.7866	Q01518
LVTASQCQQPAENK	0.9998	1688.7867	Q01518
LVTGPLVLNR	0.8597	1168.6716	Q02878
LVTYVPVTTFK	0.7844	1382.7598	P62899
LVYQEPIPTAQLVQR	0.9940	1841.9788	P25787
LYGLGELPQGFAR	0.8151	1507.7571	P31150
LYYTGEKGQNQDYR	0.9828	1849.8383	P19338
MADSGLLLK	0.9534	1062.5531	Q5SY16
MANAGPNTNGSQFFICTAK	0.9313	2143.9567	A2BFH1
MAPKPGPYVK	0.9679	1230.6583	Q01518
MAPKPGPYVK	0.8610	1230.6583	Q01518
MAPKPGPYVK	0.9982	1230.6509	Q01518
MAPKPGPYVK	0.9996	1230.6509	Q01518
MAPKPGPYVK	0.9377	1230.6509	Q01518
MGNHELYMR	0.9480	1237.5120	P15311
MIEIMEMK	0.9947	1139.5177	P40227
MIEPIDEYCVQQLK	0.9925	1880.88	P07900
MIGLPGAGK	0.8389	958.4984	Q00839
MIVNNLLKPISVEGSSK	0.9703	1972.0815	Q9HB71
MIVNNLLKPISVEGSSK	0.9999	1972.0815	Q9HB71
MLARMASEVH	0.7382	1247.5465	B3QLW4
MMQNPQILAALQER	0.9998	1729.8391	P55209
MMTPTVLYDVQELR	0.9925	1782.8433	P09525

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
MPFVTEELFQR	0.8886	1483.6918	P26640
MTEPIDEYCVQQLK	0.9885	1868.8437	Q58FF7
MTEPIDEYCVQQLK	0.9839	1868.8437	Q58FF7
MTEPIDEYCVQQLK	0.9988	1868.8362	Q58FF7
MTSSYGHVLER	0.8976	1382.6037	P54821
MTTVHAITATQK	0.8629	1416.7183	P04406
MTTVHAITATQK	0.9988	1416.7183	P04406
MVGSYGPRPEEYEFLTPVEEAPK	0.9584	2740.2842	P52566
MVLTKMKE	0.9640	1122.5929	P11021
MVPGKPMCVESFSDYPPLGR	0.9496	2382.0959	P68104
MVPGKPMCVESFSDYPPLGR	0.7906	2382.0959	P68104
MVPGKPMCVESFSDYPPLGR	0.9647	2382.0959	P68104
MVTEALKPYSSGGPR	0.9984	1707.8402	P06744
MVTPGHACTQK	0.9750	1344.6067	P04075
MVTPGHACTQK	0.8665	1344.5992	P04075
MVTPGHACTQK	0.9986	1344.5992	P04075
MVTPGHACTQK	0.9809	1344.5992	P04075
MVVTFAPVNVTTEVK	0.8720	1749.9124	P68104
MVWEGLNVVK	0.8713	1289.659	O60361
NAAGGLNPK	0.9162	956.4828	P00491
NACFEPANQMVK	0.9552	1523.6575	P68362
NACFEPANQMVK	0.9969	1523.6575	P68362
NAHIQQVGDR	0.9943	1224.5748	Q00610
NAHIQQVGDR	0.9996	1224.5748	Q00610
NAPEQACHLAK	0.9255	1353.6247	P61981
NAPPPELLEIINEDIAK	0.7474	1991.0363	Q15084
NAPPPELLEIINEDIAKR	1.0000	2147.1374	Q15084
NAPPPELLEIINEDIAKR	0.9849	2147.1374	Q15084
NAQAAALYK	0.9992	1064.5403	P40121
NAQAAALYK	0.9072	1064.5329	P40121
NAQYLLQNSVK	0.9995	1392.7149	P61978
NAQYLLQNSVK	0.8884	1392.7075	Q3T0D0
NCDLHFK	0.9971	1048.4548	Q9BUJ2
NCDLHFK	0.9299	1048.4548	Q9BUJ2
NDGAAALVLMTADAAKR	0.9577	1802.9097	P24752
NDGATILSMMDVDHQIAK	0.9974	2073.9611	P48643
NDGATILSMMDVDHQIAK	0.9989	2073.9611	P48643
NEASVLHNLK	0.8660	1239.6360	P35580
NEASVLHNLK	0.9489	1239.636	P35580

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
NFIFGQTGAGNNWAK	0.9538	1739.8168	Q9BUF5
NFTDGALVQHQEWDGK	0.9966	1959.8864	Q01469
NFTDGALVQHQEWDGK	0.9997	1959.8864	Q01469
NFVFGQSGAGNNWAK	0.9960	1711.7855	Q13885
NHHLQETSFTK	0.7314	1456.6847	P13693
NHPGQISAGYAPVLDCHTAHIACK	0.9997	2732.2700	P68104
NHPGQISAGYAPVLDCHTAHIACK	0.9915	2732.27	P68104
NIALLSDLTK	0.9782	1202.6659	P30048
NIFISERPTDVLQTVK	0.9922	1975.0527	Q15813
NIQLVTSQIDAQR	0.9672	1572.7937	Q86V81
NLFVGNLNFNK	0.9842	1394.7095	P19338
NLLKPISVEGSSK	0.9767	1514.8456	Q9HB71
NLLKPISVEGSSK	0.9933	1514.8387	Q3T168
NLNLKPGECLR	0.9863	1428.7295	P09382
NLSYSATEETLQEVFEK	0.9837	2102.9796	P19338
NLVVKPDQLIK	0.9954	1409.8394	P53396
NLVVKPDQLIK	0.9725	1409.8394	P53396
NMILDDGGDLTNLIHTK	1.0000	1984.9676	P23526
NMLNPPAEVTTK	0.8177	1429.6949	P13010
NNDIMLIK	0.8238	1091.5433	P07477
NNDPLVLR	0.8609	1027.5199	Q9UH17
NPGLAELIAEK	0.9966	1269.6717	P06737
NPIISGLYQGAGGPGPGGFGAQGPK	1.0000	2412.1975	P08107
NPIISGLYQGAGGPGPGGFGAQGPK	0.9956	2412.1975	P08107
NPNIPNEIIHALQAGR	0.9986	1843.9441	P52272
NPNTNDLFNAVGDGIVLCK	0.9996	2176.0371	P13796
NPVDILTYVAWK	0.9663	1533.7905	P00338
NSASAIGCHVVNIGAEDLK	0.9999	2069.9952	P13796
NSFVNDIFER	0.9173	1327.5945	P33778
NSFVNDIFER	0.9997	1327.5945	P33778
NSFVNDIFER	0.9754	1327.5945	P33778
NSFVNDIFER	0.8381	1327.5871	P06900
NSFVNDIFER	0.7422	1327.5871	P06900
NSPVWGADKCEELLEK	0.9739	2017.9567	O75367
NSQLPVDHILAGSFETAMR	0.9976	2173.0374	P53621
NTHADFADECPKPE	0.9939	1745.7103	P43487
NTIDTLLSVVEDHK	0.9999	1698.8577	O95373
NVDLSTVDKDQSIAPK	0.9144	1872.9581	P04844
NVDLTEFQTNLVPYPR	0.9993	1992.9617	P68362

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
NVDLTEFQTNLVPYPR	0.9992	1992.9617	P68362
NVGAGGPAPAAGAAPAGGPAPSTAAAPAEEK	0.9985	2698.3099	P05386
NVLPVFDNLMQQK	0.8634	1660.8395	P07339
NVNIGSLICNVGAGGPAPAAGAAPAGGPAPSTAAAPAEEK	1.0000	3668.8005	P05386
NVSAVDKSTGKE	0.9877	1377.6888	P11142
PDASKPEDWDER	0.9999	1559.6567	Q4VIT5
PGLHVWR	0.9807	951.4753	P40121
PLLSGLLDSPALK	0.9210	1438.8109	P49327
PPAGSAPGEHVFVK	0.9999	1507.7497	P54577
PPATQKAK	0.8787	983.5552	P50502
PPAVAPR	0.8262	794.4187	Q9H9H4
PTGTYHGDSDLQLDR	0.9983	1761.7637	Q9YHC3
QAALKNPPINTK	0.8806	1437.8092	O15511
QAAVYFEKGDYNK	0.9943	1675.7994	P31948
QAELAVILK	0.9447	1099.6389	Q14980
QALLELEMNSDLK	0.9916	1618.8024	P62081
QALLELEMNSDLK	0.9636	1618.795	A6H769
QANPILEAFGNAK	0.9762	1487.7520	P35749
QARPDDLLISTYPK	0.9508	1731.8944	P50225
QAYQEAFEISKK	0.9994	1584.7936	P31946
QCQQPAENK	0.9392	1217.5247	Q01518
QDAAIVGYK	0.9882	1079.5399	P07737
QEAFEISKK	0.9981	1222.6346	P31946
QEYDESGPSIVHR	0.9416	1586.6677	P53478
QGESITHALK	0.9999	1198.6094	Q01518
QGESITHALK	0.9810	1198.6094	Q01518
QGESITHALK	0.9998	1198.6094	Q01518
QGESITHALK	0.9993	1198.6094	Q01518
QGESITHALK	0.9877	1198.602	Q01518
QGESITHALK	0.9987	1198.602	Q01518
QGESITHALK	0.9963	1198.602	Q01518
QGESITHALK	0.9982	1198.602	Q01518
QGESITHALK	0.9891	1198.602	Q01518
QGESITHALK	0.9831	1198.602	Q01518
QGESITHALK	0.9943	1198.602	Q01518
QGESITHALK	0.7683	1198.602	Q01518
QGESITHALK	0.9301	1198.602	Q01518
QGESITHALK	0.9684	1198.602	Q01518
QGGVLPNIQAVLLPK	0.9772	1661.9617	Q96QV6

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
QGGVLPNIQAVLLPK	0.9948	1661.9617	Q96QV6
QGGVLPNIQAVLLPK	0.9384	1661.9542	P04908
QGTVIHFNNPK	0.9952	1369.6891	Q13892
QHGKVEIIANDQGNR	0.8943	1793.8921	P34931
QHLGESTVR	0.9486	1113.5315	P06576
QHNDIIR	0.9882	982.4733	Q14697
QLGNIVFK	0.9433	1033.5709	P35749
QLLLENLGNENVHR	0.7279	1735.8753	Q14974
QLLLFASK	0.9758	1034.5912	Q08211
QLTHSLGGGTGSGMGTLLISK	0.9951	2130.0892	Q13885
QLTHSLGGGTGSGMGTLLISK	0.9508	2130.0892	Q13885
QPILLELEAPLK	0.9997	1478.8496	P62136
QPILLELEAPLK	0.9976	1478.8427	Q61JR3
QPLVILEMESGASAK	1.0000	1687.8603	P49588
QPPAAPPAAPALSAADTKPGTTGSGAGSGGPGGLTSAAPAGGDKK	0.9670	4039.0399	P67809
QPSFLGMESCGIHETTFNSIMK	1.0000	2629.1763	P60709
QPSFLGMESCGIHETTFNSIMK	0.9992	2629.1763	P60709
QPTVGMNFKTPRGPV	0.9243	1743.8879	P08708
QQLDLTHLK	0.9518	1210.6458	P61221
QQNGIVPIVEPEILPDGDHDLK	0.9951	2541.2863	P04075
QSEVKPILEK	0.9773	1313.7343	P30153
QSPVDIDTHTAK	0.9999	1409.6501	P00918
QVAEVFTGHMGK	1.0000	1418.6765	P06576
QVAEVFTGHMGK	0.8758	1418.669	Q5ZLC5
QYLLTLGFK	0.8629	1197.6546	Q9BXB7
QYLLTLGFK	0.8742	1197.6546	Q9BXB7
QYLLTLGFK	0.8936	1197.6546	Q9BXB7
RDQNILLGTTYR	0.9965	1536.7797	P78527
RDTKENGKHMDL	0.9712	1586.7623	Q86XP1-3
RDTKENGKHMDL	0.8933	1586.7623	Q86XP1-3
REPIICK	0.7282	1030.5382	P48735
REPVVTLEGHTK	0.9857	1480.7786	P31146
RFDQLFDDESDPFEVLK	0.9879	2215.0222	Q8NC51
RIVAPGKGILAADE	0.9780	1524.8412	P04075
RLPLQDVYK	0.9964	1246.6822	P68104
RLPLQDVYK	0.9563	1246.6822	P68104
RLPLQDVYK	0.7985	1246.6748	P68103
RNPLIAGK	0.9675	983.5664	P62316
RPDNFVFGQSGAGNNWAK	0.9997	2079.9663	Q13885

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
RPDNFVFGQSGAGNNWAK	0.9995	2079.9663	Q13885
RPGLEGYALPR	0.9196	1315.6785	P33992
RPGLVVVHAEDGTTSK	0.9913	1780.9220	P30049
RPQNYLFGCELK	0.9561	1639.7929	P06748
RPQNYLFGCELK	0.9839	1639.7929	P06748
RPQNYLFGCELK	0.9942	1639.7857	Q3T160
RSDSENILTNYENQSR	0.9119	2012.8936	Q03001
RTVSLGAGAKDE	0.9504	1318.6629	P06748
RTVSLGAGAKDE	0.9768	1318.6629	P06748
RVHIPNDDAQFD	0.9955	1513.6698	Q16576
RVPAGNWVLIEGVDQPIVK	0.9954	2205.2058	Q15029
SAAAVLSHNR	1.0000	1112.5475	P04844
SAALIQQATTVK	0.9947	1345.7354	P32969
SAALIQQATTVK	0.9753	1345.7354	P32969
SADFPALVVK	0.8513	1161.6182	P22102
SADPEGHFETPIWIER	0.9965	1970.8911	Q14697
SAGIMDHEEAR	0.9771	1302.5411	P62244
SAGIMDHEEAR	0.9961	1302.5411	P62244
SAGTQCLISGWGNTK	0.9998	1694.776	P00760
SAGTQCLISGWGNTK	0.9999	1694.7757	P00760
SAGTQCLISGWGNTK	0.9999	1694.7757	P00760
SAGTQCLISGWGNTK	0.9938	1694.7757	P00760
SALAAATAAAAAAASAAAATAA	0.8837	1802.8837	Q99932
SALGIPSLLPFLK	0.9303	1470.8598	O75533
SALGIPSLLPFLK	0.9911	1470.8527	O75533
SALILHDDE	0.8397	1099.4934	P05386
SAPAFSLVFPFLK	0.9691	1538.8211	Q92616
SAPGPLELDLTGDLESFKK	0.9879	2160.1102	P52565
SAPKPQTSPSPK	0.9626	1367.7197	Q01518
SAPKPQTSPSPK	0.9833	1367.7197	Q01518
SAPVLAVAGLGDSNQFFR	0.9997	1935.9591	Q9UHL4
SASAIGCHVVNIGAEDLK	0.9018	1955.9523	P13796
SASAIGCHVVNIGAEDLK	0.9997	1955.9523	P13796
SASAIGCHVVNIGAEDLK	0.9998	1955.9523	P13796
SASIPGILALDLCPSDTNK	0.9949	2087.0357	Q9UMS4
SASIPGILALDLCPSDTNK	0.9226	2087.0357	Q9UMS4
SASTPVFGGILSLINEHR	0.9581	1985.0119	O14773
SATMPSDVLEVTK	0.7211	1492.7231	P60842
SATMPSDVLEVTKK	0.9995	1648.8494	P60842

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
SATMPSDVLEVTKK	0.8368	1648.8494	P60842
SCGLTHTAVVPLDLVK	0.9998	1824.9556	Q00325-2
SCVGVFQHGK	0.9991	1233.5713	P34931
SCVGVFQHGK	0.9645	1233.5713	P34931
SDNAPPPELLEIINEDIAKR	0.9927	2349.1964	Q15084
SDYPPLGR	0.9982	991.4511	P68104
SDYPPLGR	0.9501	991.4437	P68103
SEAYLVGLFEDTNLCAIHAK	1.0000	2366.1365	Q71DI3
SEIAALLVKPQK	0.9563	1439.8500	Q8NF50
SENFQTLLDAGLPQK	0.9984	1775.8767	O60506
SENFQTLLDAGLPQK	0.9993	1775.8767	O60506
SEVILPVPAFNVINGGSHAGNK	0.7623	2335.2073	P06733
SFNPYSEFILATGSADK	0.8721	1961.9159	Q09028
SFVDKDLLEPGCSVLLNHK	0.9514	2314.1780	P62191
SGGPVVCSGK	0.9481	1062.4842	P00760
SGGPVVCSGK	0.9979	1062.4842	P00760
SGGPVVCSGK	0.9972	1062.4842	P00760
SGGQHTVLLVK	0.9660	1253.6880	P18754
SGGTTMYPGIADR	0.9826	1428.6092	P62736
SGGTTMYPGIADR	0.9989	1428.6092	P62736
SGGTTMYPGIADR	0.9987	1428.6018	P53478
SGGTTMYPGIADR	0.9946	1428.6018	P53478
SGGTTMYPGIADR	0.7323	1428.6018	P53478
SGGTTMYPGIADR	0.9802	1428.6018	P53478
SGGTTMYPGIADR	0.7711	1428.6018	P53478
SGGTTMYPGIADR	0.9982	1428.6018	P53478
SGGTTMYPGIADR	0.9987	1412.6143	P62736
SGGTTMYPGIADR	0.9030	1412.6143	P62736
SGGTTMYPGIADR	0.9991	1412.6068	P53478
SGGTTMYPGIADR	0.9992	1412.6068	P53478
SGGTTMYPGIADR	0.9992	1412.6068	P53478
SGGTTMYPGIADR	0.9990	1412.6068	P53478
SGGTTMYPGIADR	0.9888	1412.6068	P53478
SGGTTMYPGIADR	0.9345	1412.6068	P53478
SGGTTMYPGIADR	0.9773	1412.6068	P53478
SGGTTMYPGIADR	0.9633	1412.6068	P53478
SGGTTMYPGIADRMQKE	0.8656	1956.8822	P62736
SGGVLPNIHPELLAK	0.9742	1659.9096	O75367
SGGVLPNIHPELLAK	0.9738	1659.9096	O75367

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
SGIPAGWMGLDCGPESSKK	0.9998	2119.9819	P00558
SGIPAGWMGLDCGPESSKK	0.9960	2119.9819	P00558
SGSPFPGSVQDPGLHVWR	0.9994	2009.9496	P40121
SGSPFPGSVQDPGLHVWR	0.9998	2009.9496	P40121
SGSPFPGSVQDPGLHVWR	0.9989	2009.9417	P40121
SGSSHQDLSQR	0.9998	1288.5544	P11908
SGSSHQDLSQR	0.9719	1288.5544	P11908
SGVTTCLR	0.9598	980.4498	Q13885
SGVTTCLR	0.7225	980.4498	Q13885
SGVTTCLR	0.8879	980.4498	Q13885
SHDASTNGLINFIK	0.9948	1631.8056	P06744
SHDGAFLAVCDASK	1.0000	1592.7041	O75083
SHPLIPDK	0.9203	1021.5345	Q9NSE4
SHTLAVDAK	0.9560	1056.5352	Q14289
SISIALIGGSR	0.9414	1160.6301	Q86UK0
SIVPALEIANAHR	0.9475	1477.7789	P10809
SIVPALEIANAHR	0.9964	1477.7717	P10809
SIVPALEIANAHR	0.9713	1477.7717	P10809
SLAGGIIGVK	0.9877	1029.5971	P61978
SLDIQCEELSDAR	0.9595	1622.692	P13489
SLFLTDLYSPEYPGPSHR	0.9993	2166.0170	Q16181
SLFLTDLYSPEYPGPSHR	0.9982	2166.017	Q16181
SLGGGTGSGMGTLLISK	0.9150	1650.8327	Q9YHC3
SLGGGTGSGMGTLLISK	0.9758	1650.8327	Q9YHC3
SLIALVNDPQPEHPLR	0.9995	1885.9798	P68036
SLIALVNDPQPEHPLR	0.9947	1885.9798	P68036
SLLAGPVAEYLK	0.8111	1375.7499	Q01518
SLLAGPVAEYLK	0.9676	1375.7425	Q01518
SLLAGPVAEYLK	0.8999	1375.7425	Q01518
SLLDKFLIK	0.9230	1219.7328	Q04917
SLLDKFLIK	0.7952	1219.7328	Q04917
SLLDKFLIK	0.9389	1219.7328	Q04917
SLPLDTLLVDVEPK	0.9969	1653.8977	P62314
SLPLDTLLVDVEPK	0.9666	1653.8977	P62314
SLPLDTLLVDVEPK	0.9398	1653.8977	P62314
SLPLITASILSK	0.9997	1357.7969	P19971
SLVDLKAELFR	0.9095	1405.7717	Q6PII3
SMANAGPNTNGSQFFICTAK	0.9998	2230.9888	A2BFH1
SMMDVDHQIAK	0.9988	1389.6169	P48643

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
SMMGQLMKPK	0.9698	1293.6395	Q9Y265
SNLLDLNPQNINK	0.9925	1597.8212	P63010
SNLNLKPGECLR	0.9998	1515.7616	P09382
SNNLCLHFNPR	0.9874	1458.65	P09382
SNPLLEAFGNAK	0.7821	1375.6884	Q12965
SNQGGLVHPK	0.9914	1151.5836	P56537
SNVLIIGELLK	0.8793	1313.7707	Q92526
SNVLIIGELLK	0.8850	1313.7707	Q92526
SPAPAAAAPAVQ	0.7793	1137.5492	P51610
SPAVHLDLLSLR	1.0000	1407.7622	P34810
SPPVAMETASTGVAAVP	0.8109	1671.7852	Q64548
SPSSSIVPAFNTGTITQVIK	0.9021	2162.1371	O43747
SQCGSLIGK	0.9208	1064.5073	Q15366
SQCQQPAENK	0.9994	1304.5567	Q01518
SQCQQPAENK	0.999	1304.5567	Q01518
SQCQQPAENK	0.9286	1304.5567	Q01518
SQEESIKPK	0.8264	1188.6138	Q9UBQ5
SQHQALLGTIR	0.9995	1310.6843	P25705
SQLLNGLK	0.9120	987.5501	Q8NCG7
SQLSAAVTALNSESNFAR	1.0000	1952.9340	O75390
SQLSAAVTALNSESNFAR	0.9513	1952.9340	O75390
SQQAYQEAFEISKK	0.9999	1799.8842	P31946
SQVLAGLMEAQK	0.9989	1389.7074	Q96EK9
SQVTTVCQALAK	0.9295	1420.7133	P50897
SSAPGPLELDLTGDLESFKK	0.9999	2247.1423	P52565
SSEGVPDLLV	0.8156	1102.522	Q0Q473
SSEPACLAEIEEDKAR	0.9771	1919.8683	P78527
SSFGPISEVVVVK	0.9673	1462.7820	P98179
SSGFSLEDPQTHSNR	0.8627	1748.7502	P08238
SSGGLSKDDIENMVK	0.9998	1722.8246	P38646
SSGGLSKDDIENMVK	0.9965	1722.8246	P38646
SSHIANVER	0.9134	1099.5158	P06396
SSIVPAFNTGTITQVIK	0.7206	1891.0203	O43747
SSLAATLLANHSLR	0.9999	1540.8109	P13489
SSLAATLLANHSLR	0.9998	1540.8109	P13489
SSLAATLLANHSLR	0.9980	1540.8037	P13489
SSLMGLFEK	0.9992	1126.5481	P50395
SSMAEVDAAMAARPHSIDGR	0.9721	2158.9636	P22626
SSMLREQILDLSK	0.9145	1650.8399	Q09328

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
SSSQELGAALAQLVAQR	0.9946	1815.9157	O95336
SSTFDAGAGIALNDHFVK	0.9856	1964.9380	P04406
STAIQELFK	0.9355	1151.5975	Q13885
STAIQELFK	0.7894	1151.59	Q9YHC3
STAIQELFKR	0.9821	1307.6986	Q13885
STAIQELFKR	0.9231	1307.6986	Q13885
STDLPLNIECFMNDKDVSGK	0.8491	2426.1246	Q92598
STSHVPEVDPGSAELQK	0.9750	1895.9013	P49327
STSLLGPPPGLLTPPVATELSQNAR	0.9982	2603.3707	Q8N163
STVFKDDDDVVIGK	0.9879	1680.8359	O15144
SVEQITAMLLTK	0.9024	1448.7697	Q92598
SVEVDGNSFEASGPSKK	1.0000	1880.8904	Q12906
SVEVDGNSFEASGPSKK	0.9927	1880.8904	Q12906
SVGIDHLALDEIK	0.9997	1524.7936	Q9UBQ7
SVHYPGEAVATR	0.8563	1373.6397	O94808
SVLNLVIVK	0.8213	1099.6753	P62753
SVLNVLHSLVDK	0.9997	1438.7932	Q9Y262
SVLTQSVK	0.9674	976.5341	P62316
SVPAVPGALGPLTITSSAVTGR	0.9408	2138.1484	O95758
SVSLVADENPFAQGALK	0.9999	1860.9370	P06396
SVSLVADENPFAQGALK	0.9945	1860.9297	P06396
SVSLVADENPFAQGALK	0.9450	1860.9297	P06396
SVSLVADENPFAQGALR	0.9914	1860.9044	P13020
SVTLHQDQLK	0.7908	1283.6622	P53602
SVVIIAAELLK	0.9977	1270.7648	P17987
SYLGGFDSSSNVLAGQLR	0.8691	1957.9282	Q6XQN6
SYPLSEGQLDQK	0.9229	1479.6919	P23141
SYVGDEAQSK	0.8796	1198.518	P53478
SYVGDEAQSK	0.9994	1198.518	P53478
SYVGDEAQSK	0.9993	1198.518	P53478
SYVGDEAQSKR	0.9980	1354.6191	P53478
SYVGDEAQSKR	0.9995	1354.6191	P53478
SYVGDEAQSKR	0.9967	1354.6191	P53478
SYVGDEAQSKR	0.8097	1354.6191	P53478
SYVGDEAQSKR	0.9919	1354.6191	P53478
SYVGDEAQSKR	0.9607	1354.6191	P53478
TAASSSSLEK	0.9267	1095.5122	P53478
TAASSSSLEK	0.8598	1095.5122	P53478
TAEILELAGNAAR	0.9973	1415.7156	P04908

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
TAHIACK	0.9615	915.4385	P68104
TAHIACK	0.9343	915.4385	P68104
TALLQPHDR	0.8171	1137.5679	P34897
TAPVNIAVIK	0.7510	1140.6655	P53602
TAQLDEELGGTPVQSR	0.9848	1787.8367	P06396
TDINLPYLTMDSSGPK	0.9244	1866.8822	P38646
TDINLPYLTMDSSGPK	0.9963	1866.8747	P38646
TDINLPYLTMDSSGPK	0.9865	1866.8747	P38646
TGLAWSKTGPVAKE	0.9973	1587.8409	Q01518
TGNVPLKVGQK	0.9258	1283.7350	Q8WUE5
THGLNEEQR	0.8291	1170.5166	O95373
THNMDVPNIK	0.9235	1283.6080	P63241
THSLGGGTGSGMGTLLISK	0.9465	1904.9414	Q13885
THSLGGGTGSGMGTLLISK	0.8868	1904.9414	Q13885
THSLGGGTGSGMGTLLISK	0.9927	1888.9465	Q13885
THSLGGGTGSGMGTLLISK	0.9678	1888.9465	Q13885
THSLGGGTGSGMGTLLISK	0.8534	1888.9465	Q13885
THSLGGGTGSGMGTLLISK	0.9999	1888.9387	Q9YHC3
THSLGGGTGSGMGTLLISK	0.9999	1888.9387	Q9YHC3
THSLGGGTGSGMGTLLISK	0.9999	1888.9387	Q9YHC3
THSLGGGTGSGMGTLLISK	0.7353	1888.9387	Q9YHC3
TISANGDKEIGNIISDAMKK	0.9180	2276.1834	P10809
TLAVNAAQDSTDLVAK	0.9947	1731.8717	Q32L40
TLAVNAAQDSTDLVAK	0.9925	1731.8717	Q32L40
TLHLLPCEVAVDGPAPVGR	0.9976	2088.0574	Q8TDP1
TLPAGPEIGPSPAPPYGLFVGGR	0.9532	2337.1906	Q8IZ83
TLPHEILEMTNK	0.9962	1540.7707	P38919
TMSGVTTCLR	0.9926	1212.5379	Q13885
TNDFLSLLEK	0.9997	1294.6557	O14929
TNLCAIHAK	0.9961	1142.558	P84227
TNLCAIHAK	0.9897	1142.558	P84227
TPAPVEKSPAK	0.9137	1267.6924	P16401
TPEIMAPILANADVQER	1.0000	1954.9570	Q16186
TPEIMAPILANADVQER	0.9976	1954.9570	Q16186
TPGVAADLSHIETK	0.9045	1553.7763	P40926
TPGVAADLSHIETK	0.9894	1553.7767	P40926
TPIEGMLSHQLK	0.9997	1468.7496	Q9UQ80
TPLLDYALEVEK	1.0000	1505.7765	P53396
TPLLPSTTGLLNDNTFAQCK	0.9999	2306.1365	O43175

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
TPLLPSTTGLLNDNTFAQCK	0.9383	2306.1365	O43175
TPRPVIVEPLEQLDDEDGLPEK	0.9863	2604.3071	P23246
TQCGSLIGK	0.8146	1078.5229	Q15365
TQPPPAPAPHATLPR	0.9838	1637.8426	P49327
TQQLHAAMADTFLEHMCR	0.8911	2246.9771	P14618
TSAGIMDHEEAR	0.9999	1403.5887	P62244
TSEGVCLAVEK	0.9896	1307.6179	P28066
TSELDMSESKTR	0.9268	1498.6721	Q86X24
TSILEYPIEPSGVLGAVATK	0.9762	2160.1466	Q96PU8
TTAIAEAWAR	0.8833	1176.5675	Q71U36
TTHELTIPNNLIGCIIGR	1.0000	2109.0789	Q15365
TTNCLAPLAK	0.9794	1203.6070	O14556
TTSAGIMDHEEAR	0.9998	1504.6364	P62244
TTSAGIMDHEEAR	0.9990	1504.6364	P62244
TTSAGIMDHEEAR	0.9997	1504.6364	P62244
TTSAGIMDHEEAR	0.9975	1504.629	Q9LX88
TTSHELTIPNDLIGCIIGR	0.9981	2197.0949	Q15366
TTVHAITATQK	0.9965	1285.6779	P04406
TTVHAITATQK	0.9963	1285.6779	P04406
TTVHAITATQK	0.8932	1285.6779	P04406
TVEGPPPKDTGIAR	0.9959	1552.7998	P14678
TVEGPPPKDTGIAR	0.8886	1552.7998	P14678
TVFTPLEYGACGLSEEK	0.9876	2015.9298	Q16881
TVLSGGTTMYPGIADR	0.9868	1741.8093	P60709
TVLSGGTTMYPGIADR	0.9975	1725.8067	P53478
TVLSGGTTMYPGIADR	0.9984	1725.8067	P53478
TVLSGGTTMYPGIADR	0.9917	1725.8067	P53478
TVPIYEGYALPHAILR	0.9212	1899.9995	P60709
TVPIYEGYALPHAILR	0.9998	1899.9995	P60709
TVPIYEGYALPHAILR	0.9969	1899.9995	P60709
TVPIYEGYALPHAILR	0.9976	1899.9917	P53478
TYGWTANMER	0.8779	1315.5403	Q58FG1
VADLAESIMK	0.9952	1191.5957	P00338
VADLQLIDFEGKK	0.9945	1618.8719	Q9Y376
VAEITNACFEPANQMVK	0.9237	2036.9448	Q71U36
VAEITNACFEPANQMVK	0.8759	2036.9448	Q71U36
VAGLVAHSDLDER	0.9344	1468.7058	O60506
VAGLVAHSDLDER	0.9984	1468.6984	O60506
VAIVDPHIKVD	0.9817	1320.7190	Q14697

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
VALAGLLAAQK	0.9524	1169.692	P23368
VALDFEQEMATAASSSSLEK	0.9998	2229.0259	P60709
VAPPGARQGQQQAGGDGKTE	0.9937	2066.9882	Q00839-2
VASNLNLKPGECLR	0.9986	1685.8671	P09382
VASNLNLKPGECLR	0.7846	1685.8597	P09382
VEQNFPAIAIHR	0.9225	1481.7527	O00148
VFFDIAVDGEPLGR	0.9987	1621.7814	Q6DTV9
VFGPDKK	0.9165	933.5072	P30041
VGLFEDTNLCAIHAK	1.0000	1802.8773	P68431
VGMGQKDSYVGDEAQSKR	0.9982	2097.9902	P62736
VHAITATQK	0.9936	1083.5825	P04406
VHAITATQK	0.9510	1083.5825	P04406
VHAITATQK	0.9848	1083.5751	P10096
VHAITATQK	0.9376	1083.5751	P10096
VHAITATQK	0.9255	1083.5751	P10096
VHAITATQK	0.9617	1083.5751	P10096
VHGALAPLAIPSAAAAAAAAGR	0.9954	2014.0860	P26599
VHGALAPLAIPSAAAAAAAAGR	0.9237	2014.0860	P26599
VHGALAPLAIPSAAAAAAAAGR	0.9967	2014.086	P26599
VHLDLLSLR	0.9999	1152.6403	P34810
VHLDLLSLR	0.9197	1152.6403	P34810
VIANPVNSTIPITAEVFKK	1.0000	2184.2306	P40926
VIPAAHPVGT	0.9226	1048.538	B1KHV0
VLAAELLR	0.9773	971.5552	P78371
VLAKPTPK	0.8963	996.6120	Q92954
VLIPTEGGDFNEFPVPEQFK	0.9999	2378.1583	Q01518
VLLGPPGAGKGTQAPRLAE	0.9995	1947.0690	P54819
VLPHILDTGAAGR	0.8570	1406.7418	Q15084
VLQPGTALFS	0.7304	1119.5638	P47224
VMGLLSNNNQALR	0.9868	1516.7568	Q13283
VMVGMGQK	0.9478	964.4622	P62736
VMVGMGQKDSYVGDEAQSK	0.9989	2171.9980	P62736
VMVGMGQKDSYVGDEAQSKR	0.9801	2328.0991	P62736
VNHPQVSALLGEEDEEALHYLTR	0.9998	2707.2990	Q01105
VNITPAEVGVLVGK	0.9984	1510.8507	P07737
VNITPAEVGVLVGKDR	0.9146	1781.9788	P07737
VPAPLPKKISSE	0.9097	1408.8078	P14317
VPGKPMCVESFSDYPPLGR	0.9989	2251.0554	P68104
VPITFQVK	0.9373	1046.5913	P05107

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
VQAFDSLLAGPVAEYLK	0.9991	1936.0017	Q01518
VQAFDSLLAGPVAEYLK	0.9969	1936.0017	Q01518
VQAFQFTDKHGE	0.9683	1521.7001	Q06830
VQEISHLIEPLANAAR	0.9999	1847.9641	Q9Y490
VQGLGENVTIESVADYFK	0.9977	2084.0214	P35637
VQSGMVVGLGTGSTTAFV	0.9204	1797.8647	Q8DJF2
VQSGSHLAAR	0.9763	1112.5475	P04040
VQSGSHLAAR	0.9057	1112.5475	P04040
VQSGSHLAAR	0.9966	1112.5475	P04040
VSAVDKSTGKE	0.9698	1263.6459	P11142
VSLGGFEITPPVVLR	0.9934	1670.9144	P06748
VSLGGFEITPPVVLR	0.9968	1670.9067	Q3T160
VSLINLAMK	0.9917	1103.6161	Q96QK1
VSNLVIEDTELK	0.9987	1474.7667	Q01518
VSSFYHAFSGAQK	0.8008	1543.7208	P12814
VSVLQLFCSSPK	0.9260	1479.7544	Q9Y678
VTASQCQQPAENK	1.0000	1575.7099	Q01518
VTASQCQQPAENK	0.9983	1575.7025	Q01518
VTASQCQQPAENK	0.9999	1575.7025	Q01518
VTASQCQQPAENK	0.9991	1575.7025	Q01518
VTASQCQQPAENK	0.9202	1575.7025	Q01518
VTASQCQQPAENK	0.9999	1575.7025	Q01518
VTASQCQQPAENK	0.9999	1575.7025	Q01518
VTKYTSAK	0.9859	1040.5654	P06899
VTKYTSSK	0.8905	1056.5604	Q96A08
VVAVLPHILDTGAAGR	0.9999	1675.9158	Q15084
VVAVLPHILDTGAAGR	0.9998	1675.9158	Q15084
VVSAAHCYK	0.9567	1149.5389	P35030
VVSAAHCYK	0.9961	1149.5389	P35030
VWNTHADFADECPKPELLAIR	0.7899	2597.2485	P43487
WDMLDLAK	0.9126	1106.5218	P49321
WIGENVSGLQR	0.9761	1345.6527	Q14019
WQGLIVPDNPPYDK	0.9828	1756.8573	P68036
WVAMAPKPGPYVK	0.9953	1586.8431	Q01518
WVAMAPKPGPYVK	0.9826	1586.8431	Q01518
WVVSAAHCYK	0.8515	1335.6182	P35030
WVVSAAHCYK	0.8042	1335.6182	P35030
YADPVSAQHAK	0.9995	1301.6152	P26599
YAGAAVDELGK	0.7825	1208.5825	P30086

Supplementary Table 1 (cont.). APRc cleavage sites identified from a tryptic peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
YAHELPK	0.9609	972.4817	P46777
YALPHAILR	0.9887	1140.6192	Q562R1
YEGYALPHAILR	0.9998	1489.7466	Q562R1
YLVPILTQTLTK	0.8965	1504.8653	Q14974
YMVGPIEEAVAK	0.9512	1421.7013	P06576
YSCVGVFQHGK	0.7349	1396.6346	P34931
YTLIVRPDNTYEVK	0.9734	1825.9362	P27797
YVELQKEEAQK	0.9996	1507.7670	Q00839
YVELQKEEAQK	0.9305	1507.7670	Q00839
YVELQKEEAQK	0.9916	1507.7597	Q00839
YVELQKEEAQK	0.9663	1507.7597	Q00839
YVLLLMGAFS	0.9045	1216.5876	P37296
YVRPLPPAAIESPAVAAPAYSR	0.9838	2383.2436	P04792
YVRPLPPAAIESPAVAAPAYSR	0.9985	2383.2436	P04792
YYLLSGAGEHLK	0.8881	1465.7353	P35579

Supplementary Table 2. APRc cleavage sites identified from a GluC peptide library using Mascot and X!Tandem. Peptides identified by LC-MS/MS spectrum-to-sequence assignment with Mascot and X!Tandem are listed with PeptideProphet probability score, calculated neutral mass and one exemplary accession number of a matching Uniprot protein entry is listed. This data was further processed and rendered non-redundant for generation of cleavage specificity profiles.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
AAAALAAAAVK	0.9913	1042.5923	Q8TAQ2
AAAGYDVEKNNSR	0.9739	1509.6960	Q02539
AAAIAYGLDK	0.8871	1107.5712	P11021
AAAIAYGLDKK	0.9997	1263.6975	P54652
AAEKLQVVGR	0.9851	1185.6618	O43175
AAGAGATHSPPTDLVWK	0.9989	1793.8849	P02545
AAGGGREHALR	0.8320	1181.5802	Q6ZRF8-3
AALKNPPINTK	0.8569	1309.7506	O15511
AALLALQHKAE	0.8604	1279.7036	Q15154
AASLLGKK	0.8823	930.5650	P49321
AATLEVERPLPMEVEK	0.9959	1926.9873	P51858
ADIETIGEILKK	0.8278	1472.8238	P61978
AEAMNYEGSPIKVTLATLK	0.9993	2179.1347	P06748
AEQLKNQIR	0.8820	1214.6519	P62873
AEQLKNQIR	0.8642	1214.6519	P62873
AFEISKK	0.8685	965.5334	P31946
AFEISKK	0.9468	965.5334	P31946
AFQLFDRTGDGK	0.8260	1469.7051	P60660
AFWIDKIK	0.9037	1163.6491	Q9UKV3
AGALVLADR	0.8253	972.5141	P49736
AGPTALLAHEIGFGSK	0.8372	1683.8732	Q99497
AGPVAEYLK	0.9993	1062.5498	Q01518
AHLDATTVLSR	0.9977	1270.6418	P06576
AIAEAWAR	0.9708	974.4722	Q71U36
AKDAFLGSFLYE	0.9669	1475.7007	P02769
AKPFVPNVHAAE	0.9961	1394.7095	Q8IYD1
AKQIVWNGPVGVFE	0.9815	1658.8569	P00558
ALALFGGEPK	0.8784	1117.5920	P49736
ALCSLHSIGK	0.9605	1200.6073	P14174
ALFAQLNQGE	0.9944	1177.5516	P40123
ALFAQLNQGE	0.9764	1177.5516	P40123
ALGWVAMAPKPGPYVK	0.9990	1827.9858	Q01518
ALGWVAMAPKPGPYVK	0.9940	1827.9858	Q01518

Supplementary Table 2 (cont.). APRc cleavage sites identified from a GluC peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
ALKLESCGVTSDNCR	0.9992	1824.8247	P13489
ALLSLAKGDRSE	0.9374	1374.7255	P04083
ALNGKEVAAQVK	0.8218	1370.7670	Q99497
ALPHAILR	0.9219	977.5558	Q562R1
ALPHAILRLD	0.9854	1205.6669	Q562R1
ALPHAILRLD	0.9975	1205.6669	Q562R1
ALPHAILRLD	0.9941	1205.6669	Q562R1
ALPHAILRLD	0.9268	1205.6669	Q562R1
ALQDMLLLK	0.9084	1175.6372	Q6P2M8-5
ALSRQLSSGVSE	0.9417	1320.6348	P04792
AQINQGESITHALK	0.9988	1624.8321	Q01518
ASGKQEPEAK	0.8777	1187.5934	Q92797
ASLAAAKK	0.8668	902.5337	P20700
ASPAPVK	0.9154	784.4157	Q2N9J7
ASQCQQPAENK	0.9975	1375.5938	Q01518
ASSSSLEKSYELPDGQVITIGNER	0.9989	2695.3089	P62736
ATIIDILTKR	0.9912	1258.7397	P04083
AVALAGLLAAQK	0.8766	1240.7291	P23368
AVLLGPPGAGKGTQAPRLAE	0.9979	2018.0987	P54819
AVLPPLPKRPALE	0.8666	1515.8925	Q9NR56
AVLVALKRAQSE	0.8888	1399.7935	P25786
AVRLLLPGE	0.9168	1054.5923	Q96A08
AVSSPPPADLCHALR	0.8942	1677.8045	P24534
AVTYTEHAK	0.9770	1134.5458	P62805
AVVTVPAYFND	0.9690	1282.5982	P11021
C[143	0.9742	2078.6941	A0JP86
CELINALYPEGQAPVKK	0.9871	2073.0717	P37802
CIAIKESAK	0.9978	1162.6168	P61158
CLAPLAKVIHD	0.9051	1351.7070	P04406
DANLQTLTEYLKK	0.9511	1679.8882	P55060
DANTIVCNSKDGGAWGTEQR	0.9984	2294.0134	P09382
DCHTAHIACK	0.9807	1327.5550	P68104
DEELNKLLGK	0.9736	1301.6979	P20671
DEGGFAPNILENKEGLELLK	0.9965	2329.1953	P06733
DEITYVELQKEEAQK	0.9991	1965.9683	Q00839
DESGPSIVHR	0.9749	1183.5370	Q9BYX7
DESTGSIAKR	0.9683	1178.5679	P04075
DFLLKPELLR	0.9396	1358.7710	O00148
DGCHAYLSKNSLDCE	0.9439	1883.7566	Q01518

Supplementary Table 2 (cont.). APRc cleavage sites identified from a GluC peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
DIAVDGEPLGR	0.9202	1228.5836	P62937
DISPQAPTHFLVIPK	0.9703	1777.9515	P49773
DKANAQAAALYK	0.9967	1406.7306	P40121
DKANAQAAALYKVSD	0.9983	1707.858	P40121
DKGLQTSQDAR	0.9522	1333.6374	P27797
DKVSHVSTGGGASLE	0.9249	1558.7297	P00558
DKYLIPNATQPESK	0.8801	1746.8940	P31946
DLFRGTLDPVE	0.9731	1348.6411	P11142
DNSSRPSQVVAETR	0.8458	1632.7604	P13639
DQAQKAEGAGDAK	0.8576	1431.6742	P05204
DQATSLRILNNGHAFNVE	0.8455	2085.998	P00918
DQLHAAVGASR	0.9841	1211.5795	P13804
DQSYKPDENEVR	0.9941	1594.7011	P31939
DRTAGIGGMNHFMLPD	0.9590	1834.7807	Q13SY1
DRTVIDYNGER	0.9855	1424.6432	P07237
DSLLAGPVAEYLK	0.9404	1490.7769	Q01518
DSLYVEKIDVGEAEPR	0.9038	1934.9373	P54577
DSYVGDEAQSKR	0.9958	1469.6535	P62736
DTFWKEFGTNIK	0.8305	1628.7987	Q58FF3
DTKPGTTGSGAGSGGPGGLTSAAPAGGDKK	0.9918	2728.3417	P67809
DTYNCDLHFK	0.9169	1427.5928	Q9BUJ2
DVVVLPGGNLGAQNLSESAAVK	0.9570	2253.1753	Q99497
DVVYALKR	0.9217	1078.5923	P62805
E[111	0.8931	1406.6544	Q05778
EAPLNPKANRE	0.9360	1353.6717	P18600
EDTNLCAIHAK	0.9762	1386.6350	P68431
EYIPHADLRLI	0.8602	1426.7283	P66648
FAEALAAHK	0.9996	1072.5453	P07237
FALLEIPK	0.8667	1045.5960	O94915
FAQINQGESITHALK	0.9998	1771.9005	Q01518
FAQINQGESITHALK	0.9972	1771.9005	Q01518
FASKPAAR	0.9668	962.5086	P23246
FCSEYRPK	0.9233	1201.5338	P09429
FFVQTCR	0.9814	1044.4599	B2RPK0
FGPLKAFNLVKD	0.9217	1491.8238	P26368
FHFEPNEYFTNEVLTK	0.9286	2129.9846	P55209
FIAIKPDGVQRGLVGE	0.9065	1813.9839	P15531
FIFIDSDHTDNQR	0.9454	1694.7437	P07237
FILFKDAASVEK	0.9954	1510.8184	Q99729

Supplementary Table 2 (cont.). APRc cleavage sites identified from a GluC peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
FINNPLAQAD	0.9741	1189.5516	P35579
FKELVYPPDYNPEGK	0.9917	1938.9516	P12956
FLQTPKIVADKD	0.8935	1517.8242	P07195
FLQTPKIVADKD	0.8676	1517.8242	P07195
FMILPVGAANFR	0.9942	1422.7230	P06733
FMVVNDAGRPK	0.9976	1348.6710	P11142
FMVVNDAGRPK	0.9928	1348.6710	P11142
FNVINGGSHAGNKLAMQE	0.9883	2001.9479	P06733
FNVINGGSHAGNKLAMQE	0.8442	2001.9479	P06733
FSAPKPQTSPSPK	0.9948	1514.7881	Q01518
FSQIVRVLTEDE	0.9906	1522.7337	P14324
FSTPLLLGKK	0.8938	1246.7437	P40926
FTTNLTEEEEKSK	0.9321	1698.8100	P35579
FVALSTNTTKVKE	0.9865	1580.8562	P06744
FVTFCTK	0.9503	1017.4742	O60506
FVTFDDHDPVDK	0.9668	1549.6838	P22626
FYAPWCGHCQR	0.8242	1568.6190	Q15084
FYIITNKLKE	0.9466	1411.7863	Q9NTJ3
FYVNGLTLGGQK	0.9986	1411.7248	P07737
FYVNGLTLGGQK	0.9883	1411.7248	P07737
FYVNGLTLGGQKCSVIRD	0.998	2142.068	P07737
GAFQHVGK	0.9111	958.4773	P26641
GAGLMGAGIAQVSVDKGLK	0.9604	1915.0349	P40939
GAPFLKEGASEEEIR	0.9123	1747.8529	P23141
GCHAYLSKNSLDCE	0.9738	1768.7297	Q01518
GFTLPHAILRLD	0.9621	1439.7597	P0C542
GFTLPHAILRLD	0.8971	1439.7597	P0C542
GGSHAGNKLAMQE	0.9843	1430.6360	P06733
GGSHAGNKLAMQE	0.9482	1414.6411	P06733
GGTTMYPGIADRMQKE	0.9948	1869.8501	P62736
GGTTMYPGIADRMQKE	0.9240	1869.8501	P62736
GGTTMYPGIADRMQKE	0.9773	1869.8427	P18600
GHSLGTGVATNLVR	0.8227	1468.7535	Q8N2K0
GHVLAAGCGQNPVR	0.9947	1522.7211	Q9BWD1
GIHETTFNSIMK	0.9341	1492.7132	Q562R1
GILTLKYPIE	0.9909	1261.7070	P62736
GIMNSFVNDIFERIAGE	0.9345	1998.9187	P33778
GINLVQAKKLVE	0.9807	1454.8609	P52815
GKEILVGDVGQTVDDPYATFVK	0.9997	2494.2744	P23528

Supplementary Table 2 (cont.). APRc cleavage sites identified from a GluC peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
GKVLPGVDALSNI	0.9048	1397.7597	P00558
GKVLPGVDALSNI	0.9670	1397.7597	P00558
GLAWSKTGPVAKE	0.9980	1486.7932	Q01518
GLTLGGQKCSVIRD	0.9841	1618.8249	P07737
GLTSVINQKLKDDE	0.9579	1702.8817	P07195
GLTSVINQKLKDDE	0.9769	1702.8817	P07195
GLTSVINQKLKDDE	0.9201	1702.8817	P07195
GLTSVINQKLKDDE	0.9671	1702.8817	P07195
GPIKTTE	0.9347	860.4318	P24487
GQKDSYVGDEAQSK	0.9999	1654.7587	P62736
GQLLTSSNYDDDEKK	0.9563	1855.8588	P11388
GQSGAGNNWAK	0.9782	1204.5373	Q13885
GTLLKPNMVTPGHACTQK	0.9931	2096.0659	P04075
GTLLKPNMVTPGHACTQK	0.9959	2096.0659	P04075
GVDLLADAVAVTMGPK	0.9083	1671.8654	P10809
GVGILALIDALRDNE	0.9607	1655.8557	Q9NYL9
GVLPNIQAVLLPK	0.9213	1476.8816	Q96QV6
GVLPNIQAVLLPKKTE	0.9988	1863.0982	Q96QV6
GVMVGMGQKDSYVGDE	0.9937	1786.758	P18600
GVPLSDPVPDPE	0.8852	1308.5912	A1T700
GVPMPDKYSLEPVAVELK	0.8538	2115.1074	P00558
GVSLAVCK	0.9001	948.4851	P06733
GVSLKTLHPD	0.8475	1181.6193	P00338
GYALPHAILR	0.8288	1197.6406	Q562R1
GYALPHAILR	0.9661	1197.6406	Q562R1
GYALPHAILR	0.9923	1197.6406	Q562R1
GYALPHAILR	0.9258	1197.6406	Q562R1
GYALPHAILR	0.9699	1197.6406	Q562R1
GYALPHAILR	0.9990	1197.6406	Q562R1
GYNYTGMGNSTNKK	0.9927	1677.7569	Q08211
HEKYDNSLK	0.8692	1276.6200	P04406
HGIQPDGQMPSDK	0.9991	1524.6779	Q71U36
HLATGDMLR	0.9932	1100.5185	P54819
HLQLAIRNDEE	0.9959	1424.6796	Q96QV6
HLQLAIRNDEE	0.9172	1424.6796	Q96QV6
HLQLAVRNDEE	0.9905	1410.6640	Q8IUE6
HSAVSLDPIKSFE	0.9687	1544.7623	Q9Y3F4
HVVVPVNPK	0.8494	1103.6240	Q92499
IAALVIDNGSGMCK	0.9985	1579.7486	P63261

Supplementary Table 2 (cont.). APRc cleavage sites identified from a GluC peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
IAALVIDNGSGMCK	0.9918	1579.7486	P63261
IAALVIDNGSGMCK	0.9999	1563.7537	P63261
IAALVIDNGSGMCK	0.9997	1563.7537	P63261
IAALVIDNGSGMCK	0.9920	1563.7537	P63261
IAALVIDNGSGMCK	0.9920	1563.7537	P63261
IAALVIDNGSGMCK	0.9998	1563.7537	P63261
IAALVIDNGSGMCK	0.9993	1563.7537	P63261
IAAQYSGAQVR	0.9996	1250.6156	P26641
IAEAYLGK	0.9981	979.5126	P11142
IAMATVTALR	0.9983	1133.6015	P04075
IANLFNRYPALHKPE	0.9901	1897.9951	P13796
IAPALVSKKLNVTE	0.9967	1625.9504	P06733
IAPALVSKKLNVTE	0.9985	1625.9504	P06733
IAPALVSKKLNVTE	0.9982	1625.9504	P06733
IAPALVSKKLNVTE	0.9988	1625.9504	P06733
IAPALVSKKLNVTE	0.9988	1625.9504	P06733
IAPALVSKKLNVTE	0.9926	1625.9504	P06733
IAPALVSKKLNVTE	0.9962	1625.9504	P06733
IAPALVSKKLNVTE	0.9938	1625.9504	P06733
IAPALVSKKLNVTE	0.9919	1625.9504	P06733
IAPALVSKKLNVTE	0.9985	1625.9504	P06733
IAPALVSKKLNVTE	0.9715	1625.9504	P06733
IAPALVSKKLNVTE	0.9799	1625.9504	P06733
IAPALVSKKLNVTE	0.9858	1625.9504	P06733
IAPALVSKKLNVTE	0.9326	1625.9504	P06733
IAQGGVLPNIQAVLLPKKTE	0.9916	2232.2994	Q96QV6
IAQVDPKK	0.9416	1041.5971	Q9Y2B0
IASGGVLPNIHPELLAK	0.9903	1844.0308	O75367
IFIDSDHTDNQR	0.9933	1547.6753	P07237
IFIDSDHTDNQRILE	0.9106	1902.886	P07237
IFMAIAK	0.9043	908.4942	P20339
IGAIAIGDLVK	0.9975	1184.6917	P78371
IGGIGTVPVGRVE	0.9987	1340.7201	P68104
IGLAKDDQLK	0.8252	1243.6924	P41567
IGNLNTLVVKKSDVE	0.9871	1771.9832	O60812
IITTEKTSK	0.9421	1163.6550	P40939
ILFLDPSGKVHPE	0.9712	1566.8194	O95881
ILLVQPTKRPE	0.9625	1408.8190	P84090
ILTAFQK	0.8309	935.5228	P30040

Supplementary Table 2 (cont.). APRc cleavage sites identified from a GluC peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
ILTHGIFSGPAISR	0.9912	1555.8259	P60891
ILTLKYPIE	0.9596	1204.6855	P62736
ILTLKYPIE	0.9153	1204.6855	P62736
ILTLKYPIE	0.9856	1204.6855	P62736
ILTLKYPIE	0.9931	1204.6855	P62736
ILTLKYPIE	0.9698	1204.6855	P62736
ILTLKYPIE	0.9721	1204.6855	P62736
ILTLKYPIE	0.9762	1204.6855	P62736
IMFGPDKCGE	0.9932	1268.5318	P27824
INAISKK	0.8596	916.5494	P61158
INPDHPIVETLR	0.9799	1490.7630	P08238
IQAVLLPKKTE	0.9961	1382.8285	Q96QV6
IQAVLLPKKTE	0.9959	1382.8285	Q96QV6
IQAVLLPKKTE	0.9987	1382.8285	Q96QV6
IQAVLLPKKTE	0.9980	1382.8285	Q96QV6
IQAVLLPKKTE	0.9991	1382.8285	Q96QV6
IQAVLLPKKTE	0.9399	1382.8285	Q96QV6
IQAVLLPKKTE	0.9575	1382.8285	Q96QV6
IQAVLLPKKTE	0.9968	1382.8285	Q96QV6
IQAVLLPKKTE	0.9955	1382.8285	Q96QV6
IQAVLLPKKTE	0.8761	1382.8285	Q96QV6
IQAVLLPKKTE	0.9907	1382.8285	Q96QV6
IQAVLLPKKTE	0.9986	1382.8285	Q96QV6
IQAVLLPKKTE	0.9961	1382.8285	Q96QV6
IQAVLLPKKTE	0.9952	1382.8285	Q96QV6
IQAVLLPKKTE	0.9925	1382.8285	Q96QV6
IQAVLLPKKTE	0.9940	1382.8285	Q96QV6
IQAVLLPKKTE	0.9908	1382.8285	Q96QV6
IQAVLLPKKTE	0.9273	1382.8285	Q96QV6
IQAVLLPKKTE	0.9236	1382.8285	Q96QV6
IQGITKPAIR	0.9929	1211.7138	P62805
IQGLTTAHEQFK	0.9788	1487.7521	P12814
IQNAPEQACHLAK	0.9862	1594.7674	P61981
IRNDEELNK	0.9681	1245.6101	Q96QV6
ISADIETIGEILKK	0.8922	1672.9399	P61978
ISFGTTKDK	0.8673	1139.5975	P48643
ISHLIEPLANAAR	0.9990	1491.7946	Q9Y490
ISLPIHPMITNVAK	0.9714	1648.9123	Q14204
ISPYFINTSKGQKCE	0.9990	1914.9298	P10809

Supplementary Table 2 (cont.). APRc cleavage sites identified from a GluC peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
ISPYFINTSKGQKCE	0.9980	1914.9298	P10809
ITYTDEEPVKK	0.9341	1465.7453	O14979
ITYTDEEPVKKLLE	0.9168	1820.9560	O14979
IVKLPLLPHE	0.9352	1273.7546	O43681
IVTNWDDME	0.8470	1225.4709	P60709
IYGMEGIPEKDMDER	0.9813	1897.8338	O43670
IYTNYEAGKDDYVK	0.9566	1821.8573	P09211
KACANPAAGSVILLE	0.9874	1628.8267	P00558
KACANPAAGSVILLE	0.9028	1628.8267	P00558
KACQSIYPLHD	0.8469	1446.664	Q801S3
KACSLAKTAFDE	0.9332	1483.7057	P68250
KACSLAKTAFDE	0.9819	1483.7057	P68250
KADGIVSKNF	0.8425	1221.6506	P63220
KAGAAPYVQAFD	0.8415	1352.6437	Q01518
KAPLDIPVPDPVKE	0.9922	1660.9188	Q06323
KAPNLKILNLSGNE	0.9434	1653.9202	Q9UBU9
KDDAMLLK	0.9249	1076.5688	P10809
KDFKAAID	0.9290	1050.5498	Q155Q3
KDGLILTSRGPGTSFE	0.9306	1792.9037	Q5E946
KDSPSVWAAVPGK	0.9997	1484.7776	P07737
KDSTLIMQLLR	0.9919	1432.7860	P31946
KFLIPNASQAE	0.9894	1332.6752	P63103
KFLIPNASQAE	0.9947	1332.6752	P63103
KGTVQQADE	0.8771	1090.5043	P32969
KHLIPAANTGE	0.9823	1265.6442	P62261
KHTGPNSPDTAND	0.9990	1468.6331	P31943
KHTGPNSPDTAND	0.9993	1468.6331	P31943
KKISSIQSIVPALE	0.9132	1655.9537	P10809
KLAPVPFFSLLQYE	0.8471	1766.9317	P07741
KLCYVALDFEQE	0.9410	1629.7427	P18600
KLFIGGLSFE	0.9194	1225.6421	Q32P51
KLIAPVAEEE	0.8401	1213.6267	P07195
KLRIYFLE	0.8590	1196.6705	Q96SN8
KLSDLLAPISE	0.9043	1300.6957	Q01518
KMINLSVPDTIDE	0.9941	1589.7687	P13796
KMINLSVPDTIDE	0.9604	1589.7687	P13796
KMSVQPTVSLGGFE	0.9106	1594.7813	P06748
KMSVQPTVSLGGFE	0.9985	1594.7737	Q3T160
KMVADGVEP	0.9440	1076.4886	A0M380

Supplementary Table 2 (cont.). APRc cleavage sites identified from a GluC peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
KNLSDLIDLVPSLCE	0.9525	1830.9111	P79136
KNNQITNNQR	0.9963	1344.6646	P00558
KQDRTLTIVD	0.8987	1303.6884	Q58FG1
KQGQDNLSSVKE	0.9905	1475.7368	P30040
KQGQDNLSSVKE	0.9745	1475.7368	P30040
KSPLLQLPHIEE	0.8584	1518.8194	Q9UGP8
KTVQLRNGNLQYD	0.9863	1663.8430	P06396
KVCNPIITK	0.9752	1215.6798	P11142
KVLSLLALVKPE	0.9380	1452.8997	Q9UL46
KYDPSLKPLSVSYD	0.9152	1754.8879	P00918
KYTLPPGVDPTQVSSSLSPE	0.9964	2217.0953	P04792
LAGHQTSAESWGTGR	0.9727	1644.7393	P36578
LAGPTNAIFK	0.8435	1146.6185	Q15366
LAGPVAEYLK	0.9988	1175.6338	Q01518
LAGPVAEYLK	0.9995	1175.6338	Q01518
LAIIDPGDSDIIR	0.8760	1484.7623	P62888
LAIVEALNGKEVAAQVK	0.9999	1896.0832	Q99497
LALIDKQE	0.9176	1044.5603	Q9UFN0
LALLDGSNVVFK	0.9958	1390.7608	O15212
LAPLAKVIHD	0.9088	1191.6764	P04406
LAPSTMKIKIIAPPE	0.9895	1752.0007	P62736
LAQVLAQERPK	0.9990	1367.7673	P49327
LATATGAK	0.8390	847.4552	Q9BZH6
LAVDAVIAELK	0.9986	1256.7128	P10809
LAVDAVIAELKK	0.9995	1412.8391	P10809
LAWSKTGPVAKE	0.9912	1429.7717	Q01518
LAWSKTGPVAKE	0.9891	1429.7717	Q01518
LCAIHAK	0.9092	927.4748	P68431
LFADKVPK	0.8575	1060.6069	P62937
LFAEFGTLKK	0.9604	1296.7230	Q86V81
LFAEFGTLKK	0.8950	1296.7230	Q86V81
LFHQQGTPR	0.9887	1170.5682	P20700
LFLPEEYPMAAPK	0.9695	1620.8010	P61088
LGAYCGYSAVR	0.9461	1303.5767	P21964
LGGPEAAKSDETAAK	0.9994	1587.7892	P04792
LGIILAHTNLR	0.9995	1307.7462	P31939
LGIPFAKPPLGPLR	0.9231	1590.9398	P23141
LGPKPEVAQQTR	0.8426	1438.7680	P53621
LGPLVSKVKE	0.8785	1212.7230	Q86VP6

Supplementary Table 2 (cont.). APRc cleavage sites identified from a GluC peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
LIINSLYKNKE	0.9706	1477.8292	P14625
LIINSLYKNKE	0.9920	1477.8292	P14625
LIINSLYKNKE	0.9772	1477.8292	P14625
LIINTFYSNKE	0.9733	1456.735	Q58FF8
LIQTADQLR	0.9734	1144.5988	P18031
LISVYSEKGESSGK	0.9402	1626.8253	P36578
LIVLEGVDR	0.9786	1100.5978	P23919
LIVPDNPPYDKGAFRIE	0.9091	2059.0527	P68036
LIYTNYEAGKDDYVK	0.9994	1934.9414	P09211
LKAPLDIPVPDPVKE	0.9976	1774.0029	Q06323
LKEDQTEYLEER	0.9823	1667.7790	Q58FF7
LKKAGGANYDAQTE	0.9917	1608.7817	Q2HJ57
LKKAGGANYDAQTE	0.9743	1608.7817	Q2HJ57
LKQEVISTSSK	0.8889	1362.7507	P63244
LLAAEFLK	0.9989	1019.5803	Q99832
LLAKNLPYKVTQDE	0.8899	1774.9617	P19338
LLALVKPE	0.8901	997.5960	Q9UL46
LLAYTLGVK	0.9190	1092.6331	P68104
LLDKYLIPNATQPESK	0.9954	1973.0621	P31946
LLIGPRGNTLKNIE	0.8990	1652.9362	Q15637
LLKQGQDNLSSVKE	0.9443	1701.9049	P30040
LLKQGQDNLSSVKE	0.8636	1701.9049	P30040
LLPAIVHINHQPFLE	0.9191	1827.9784	P17844
LLTSFGPLK	0.9887	1090.6175	P26368
LLVVTDPRADHQPLTE	0.9264	1890.9588	P08865
LMTPAACPEPPPEAPTEDDHDEL	0.9787	2619.0893	P14314
LNILTAFQKKGAE	0.9967	1575.8773	P30040
LNMLSLK	0.9704	933.5105	Q9Y617
LNVVDIAGLVK	0.9988	1255.7288	Q9NTK5
LQANCYEEVKDR	0.9711	1639.7412	P23528
LQGIPVLVLGNKR	0.9884	1521.9143	Q96BM9
LQKYPPPLIPPRGE	0.8925	1719.9460	P25098
LQLAIRNDEE	0.9625	1287.6207	Q96QV6
LQLFRGDTVLLK	0.9998	1517.8718	P55072
LQTVAKNKDQGTYE	0.9453	1737.8686	P60660
LQTVAKNKDQGTYE	0.9641	1737.8686	P60660
LSNLKAPLDIPVPDPVKE	0.9843	2088.1619	Q06323
LSQLQKQLAAKE	0.8444	1499.8459	P02545
LVEAIVLPMNHK	0.9434	1478.8067	P17980

Supplementary Table 2 (cont.). APRc cleavage sites identified from a GluC peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
LVFLPFADDKR	0.9643	1435.7612	P12956
LVTASQCQQPAENK	0.9662	1688.7940	Q01518
LVVLLQANRDPDAGIDE	0.9957	1924.9567	P08758
LVVLLQANRDPDAGIDE	0.9907	1924.9567	P08758
LVVLLQANRDPDAGIDE	0.9874	1924.9567	P08758
LVVLLQANRDPDAGIDE	0.9950	1924.9568	P08758
LVWVPSDKSGFEPASLKE	0.9955	2132.0942	P35579
LVYQEPIPTAQLVQR	0.9957	1841.9788	P25787
LYCLEHGIQPDGQMPSDK	0.9412	2202.9826	Q71U36
LYTLIVRPDNTYE	0.9211	1683.8256	P27797
LYYTGEKGQNQDYR	0.9834	1849.8383	P19338
MANAGPNTNGSQFFICTAK	0.9489	2143.9567	A2BFH1
MAPKPGPYVKE	0.9953	1359.6937	Q01518
MAPKPGPYVKE	0.9954	1359.6937	Q01518
MATAASSSSLEK	0.9927	1313.5921	P62736
MATAASSSSLEK	0.9515	1297.5972	P62736
MATAASSSSLEK	0.9983	1297.5972	P62736
MATAASSSSLEK	0.9918	1297.5972	P62736
MGTYATQSALSSSRPTK	0.9952	1900.9101	P53618
MLMAHAVTQLANR	0.9993	1542.7547	P78371
MQIQHPTASLIAK	0.9044	1552.8184	Q92526
MQKLDAQVK	0.9319	1203.6434	P04843
MRPGVACSVSQAQKDE	0.9679	1877.8512	P32969
MRPGVACSVSQAQKDE	0.9683	1877.8512	P32969
MSHLGRPDGVPMPD	0.9778	1595.6973	P00558
MVPGKPMCVESFSDYPPLGR	0.9712	2382.0959	P68104
NDGAAALVLMTADAAKR	0.9671	1802.9097	P24752
NDGATILSMMDVDHQIAK	0.9993	2073.9611	P48643
NGFLSPDKLSLLEK	0.9845	1703.9246	P31689
NHIIDGVK	0.9737	1010.5297	Q9GZT3
NSFVNDIFER	0.9758	1327.5945	P33778
NSKDGGAWGTEQRE	0.9945	1649.7182	P09382
NTAVSQLTKAKE	0.9970	1432.7674	Q9NTJ3
NTHADFADECPKPE	0.9808	1745.7103	P43487
NVDLSTVDKDQSIAPK	0.9395	1872.9581	P04844
NVLRQTGNNE	0.8449	1231.5693	Q9BYX4
NVPLPNTLPLPKRE	0.9946	1702.9518	Q9Y520
NVSAVDKSTGKE	0.9855	1377.6888	P11142
NVSAVDKSTGKE	0.8845	1377.6888	P11142

Supplementary Table 2 (cont.). APRc cleavage sites identified from a GluC peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
QAALKNPPINTK	0.9139	1437.8092	O15511
QARPDDLLISTYPK	0.9506	1731.8944	P50225
QGESITHALK	0.9994	1198.6094	Q01518
QGGVLPNIQAVLLPK	0.9966	1661.9617	Q96QV6
QGGVLPNIQAVLLPKKTE	0.9974	2048.1782	Q96QV6
QGLIVPDNPPYDKGAFRIE	0.9551	2244.1327	P68036
QHGKVEIIANDQGNR	0.9121	1793.8921	P34931
QIDNPDYKGTWIHPE	0.9627	1927.8853	P27797
QYLLTLGFK	0.9079	1197.6546	Q9BXB7
QYLLTLGFK	0.9159	1197.6546	Q9BXB7
QYLLTLGFK	0.9251	1197.6546	Q9BXB7
RAQPVQVAE	0.9954	1084.5413	P06396
RDQNILLGTTYR	0.9967	1536.7797	P78527
REVPCPPGTE	0.8986	1228.5295	Q9Y4B4
RIVAPGKGILAADE	0.8499	1524.8412	P04075
RIVAPGKGILAADE	0.9391	1524.8412	P04075
RIVILGPE	0.9135	983.5552	Q5T089
RKAEPEGLR	0.8924	1170.6257	Q9GZX7
RLCYVALDFEQE	0.9893	1629.7171	P43239
RLCYVALDFEQE	0.9859	1629.7171	P43239
RLCYVALDFEQE	0.9845	1629.7171	P43239
RLCYVALDFEQE	0.9592	1629.7171	P43239
RNPLIAGK	0.9790	983.5664	P62316
RPDNFVFGQSGAGNNWAK	0.9995	2079.9663	Q13885
RPGLEGYALPR	0.9260	1315.6785	P33992
RRLPLPKP	0.9056	1091.6716	Q6IE36
RSYELPDGQVITIGNE	0.9910	1877.8833	Q8BFZ3
RSYELPDGQVITIGNE	0.9957	1877.8833	Q8BFZ3
RSYELPDGQVITIGNE	0.9958	1877.8833	Q8BFZ3
RSYELPDGQVITIGNE	0.9958	1877.8833	Q8BFZ3
RVHIPNDDAQFD	0.9927	1513.6698	Q16576
RVHIPNDDAQFD	0.9248	1513.6698	Q16576
SAALIQQATTVK	0.9840	1345.7354	P32969
SAGIMDHEEAR	0.9964	1302.5411	P62244
SAIVILRPTKA	0.9582	1283.7639	Q5JFZ4
SALFAQLNQGE	0.9963	1264.5836	P40123
SALILHDDE	0.9627	1099.4934	P05386
SAPKPQTSPSPK	0.9875	1367.7197	Q01518
SAQLSQLQKQLAAKE	0.9953	1785.9663	P02545

Supplementary Table 2 (cont.). APRc cleavage sites identified from a GluC peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
SATMPSDVLEVTKK	0.8444	1648.8494	P60842
SAVPPGADKKAE	0.9062	1312.6701	Q3T0F4
SCVGVFQHGKVE	0.9471	1461.6823	P34931
SDNLKFPDLGLKLI	0.9125	1715.9536	Q57690
SFTLRQQLQTTRQE	0.9895	1822.8997	Q08E38
SGGGVAMIGVGE	0.8641	1120.4897	Q58039
SGGTTMYPGIADR	0.9304	1412.6143	P62736
SGGTTMYPGIADRMQKE	0.9848	1956.8822	P62736
SGGTTMYPGIADRMQKE	0.9972	1956.8747	P18600
SGGVTIPP	0.8890	814.3899	Q5JH10
SGSSHQDLSQR	0.9733	1288.5544	P11908
SGVTTCLR	0.9232	980.4498	Q13885
SILGTTLKDE	0.9704	1191.6135	O75083
SIQALGWVAMAPKPGPYVK	0.9990	2156.1604	Q01518
SLIALVNDPQPEHPLRADLAEE	0.9280	2514.2502	P68036
SLIINTFYSNKE	0.9769	1543.7597	Q76LV2
SLIINTFYSNKE	0.8679	1543.7597	Q76LV2
SLIINTFYSNKE	0.8433	1543.7597	Q76LV2
SLLDKFLIK	0.9482	1219.7328	Q04917
SLLLFEAMRK	0.9485	1322.7168	P47897
SLPLDTLLVDVEPK	0.9427	1653.8977	P62314
SNVLIIGELLK	0.8987	1313.7707	Q92526
SPIMAKPR	0.8343	1014.5432	Q9NP61
SPLVSRLTLYD	0.9618	1350.6857	Q32LG3
SPNSKVNTLSKE	0.9226	1446.7466	P40939
SQCQQPAENK	0.9376	1304.5567	Q01518
SQLQDTQELLQEENRQK	0.9995	2202.0665	P35579
SRGFGFVLFKE	0.9973	1401.7117	Q14103
SSEPACLAEIEEDKAR	0.9841	1919.8683	P78527
SSFYVNGLTLGGQKCSVIRD	0.9529	2316.1321	P07737
SSGFSLEDPQTHSNR	0.8535	1748.7502	P08238
SSMAEVDAAMAARPHSIDGR	0.9767	2158.9636	P22626
STGALSLKKVPE	0.8539	1372.7714	P09622
STGLSLEQVKK	0.9371	1332.7401	P16615
STRIIYGGSVTGATCKE	0.9962	1914.9258	P60174
STRIIYGGSVTGATCKE	0.9891	1914.9258	P60174
SVEVDGNSFEASGPSKK	0.9950	1880.8904	Q12906
SVLISLKQAPLVH	0.9986	1519.8797	P04973
SVLISLKQAPLVH	0.9814	1519.8797	P04973

Supplementary Table 2 (cont.). APRc cleavage sites identified from a GluC peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
SVLLPLVAKE	0.8439	1183.6964	P0C024
SVLVDAFSHVAR	0.9999	1387.6996	Q9NQG5
SVSLVADENPFAQGALRSE	0.9978	2076.9787	Q3SX14
SVYIKGFPTDATLDDIKE	0.9937	2155.0837	P05455
TAEAYLGKK	0.9108	1123.6025	P11021
TAGIQRIPLPPPPAPE	0.9454	1740.9311	Q07666
TAHIACK	0.9547	915.4385	P68104
TALLSSGFSLEDPQTHSNR	0.9996	2147.0031	P08238
TDINLPYLTMDSSGPK	0.9237	1866.8822	P38646
TGLAWSKTGPVAKE	0.9948	1587.8409	Q01518
TGLAWSKTGPVAKE	0.9515	1587.8409	Q01518
THSLGGGTGSGMGTLLISK	0.9327	1904.9414	Q13885
THSLGGGTGSGMGTLLISK	0.8932	1888.9465	Q13885
TILRPLNVEPPLTDLQK	0.9298	2062.1575	Q99459
TKVVAPTISSPVCQE	0.9985	1730.8661	Q9Y490
TLHLLPCEVAVDGPAPVGR	0.9979	2088.0574	Q8TDP1
TLLAKNLPYKVTQDE	0.9956	1876.0094	P19338
TPAPVEKSPAK	0.9201	1267.6924	P16401
TPLKPSPLPVIPDTIKE	0.9133	1988.1346	Q7Z6Z7
TPLLPSTTGLLND	0.9315	1428.7249	O43175
TPLSKLMKAYCE	0.9343	1583.7839	P61956
TQDKLYQPEYQEVSTEEQREEISGK	0.9992	3157.4839	Q9Y4L1
TRKYTLPPGVDPTQVSSSLSPE	0.9049	2474.2441	P04792
TSIANLPKLNKLKKLE	0.8991	2009.2401	P39687
TSLYTQDR	0.9701	1070.4781	P50897
TTAIAEAWAR	0.8972	1176.5675	Q71U36
TTFNSIMK	0.9936	1056.5062	Q562R1
TTGLAWSKTGPVAKE	0.8693	1688.8886	Q01518
TTSAGIMDHEEAR	0.9998	1504.6364	P62244
TTVHAITATQK	0.9219	1285.6779	P04406
TVEGPPPKDTGIAR	0.8885	1552.7998	P14678
VAGLAGKDPVQCSRD	0.9072	1687.8100	Q99497
VAKLGNREDPLPQDSFE	0.9355	2029.9857	P57737
VAPISDIIAIK	0.8800	1254.7335	P13804
VAVLPHILD	0.9575	1063.5814	Q15084
VENGGSLGSKK	0.8533	1218.6356	P14618
VEPSDTIENVKAK	0.8627	1572.8147	P62987
VFFFGTHE	0.9667	1070.4536	Q9XSK7
VGLLIGPRGNTLKNIE	0.8945	1809.026	Q15637

Supplementary Table 2 (cont.). APRc cleavage sites identified from a GluC peptide library using Mascot and X!Tandem.

Identified Peptides (prime sequence)	PeptideProphet probability	Neutral peptide mass (Da)	Exemplary protein ID
VGMGQKDSYVGDEAQSKR	0.9979	2097.9902	P62736
VHAITATQK	0.9656	1083.5825	P04406
VHAITATQKTVD	0.9856	1398.7255	P04406
VIILNHPGQISAGYAPVLD	0.9170	2064.0717	P68103
VILIDPFHK	0.9930	1196.6706	P61313
VIVVSVKEAIPGGKVKKG	0.9207	2007.2534	A8GPE0
VLAAELLR	0.9794	971.5552	P78371
VLPKLFE	0.8768	960.5432	Q14008
VLPNIQAVLLPKKTE	0.8400	1806.0767	Q96QV6
VLPNIQAVLLPKKTE	0.9563	1806.0767	Q96QV6
VMVGMGQKDSYVGDEAQSK	0.9994	2171.9980	P62736
VMVGMGQKDSYVGDEAQSKR	0.9808	2328.0991	P62736
VNITPAEVGVLVGKDR	0.9327	1781.9788	P07737
VPIILVGNKK	0.8591	1223.7754	P61586
VQAFQFTDKHGE	0.9664	1521.7001	Q06830
VQAFQFTDKHGE	0.9093	1521.7001	Q06830
VQALDDTERGSGGFGSTGKN	0.9997	2110.9668	P33316
VQSGSHLAAR	0.9976	1112.5475	P04040
VQSGSHLAARE	0.9554	1241.5901	P04040
VSTYIKK	0.9312	981.5647	P68104
VTIVNILTNR	0.9849	1229.6880	P07355
VVAVHPGGDTVAIGGVDGNVR	0.9378	2076.0501	O75083
VVAVLPHILDTGAAGR	0.9999	1675.9158	Q15084
WIVLKEPISVSSE	0.8925	1601.8453	P00918
WVAMAPKPGPYVK	0.9879	1586.8431	Q01518
YGKIDTIEIITDR	0.9996	1651.8569	P22626
YHQVIQQMEQK	0.9185	1546.7350	P80303
YQEVSTEEQREEISGK	0.9983	2026.9231	Q9Y4L1
YSCVGVFQHGKVE	0.8550	1624.7456	P34931
YSCVGVFQHGKVE	0.9617	1624.7456	P34931
YVELQKEEAQK	0.9304	1507.7670	Q00839
YVTIIDAPGHRD	0.8654	1443.6895	P68104