PROTEINS: Structure, Function, and Genetics 46:355–367 (2002)
DOI 10.1002/prot.10060

Classification of the Caspase–Hemoglobinase Fold: Detection of New Families and Implications for the Origin of the Eukaryotic Separins

L. Aravind[†] **and Eugene V. Koonin**
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

ABSTRACT **A comprehensive sequence and structural comparative analysis of the caspase–hemoglobinase protein fold resulted in the delineation of the minimal structural core of the protease domain and the identification of numerous, previously undetected members, including a new protease family typified by the HetF protein from the cyanobacterium *Nostoc*. The first bacterial homologs of legumains and hemoglobinases were also identified. Most proteins containing this fold are known or predicted to be active proteases, but multiple, independent inactivations were noticed in nearly all lineages. Together with the tendency of caspase-related proteases to form intramolecular or intermolecular dimers, this suggests a widespread regulatory role for the inactive forms. A classification of the caspase–hemoglobinase fold was developed to reflect the inferred evolutionary relationships between the constituent protein families. Proteins containing this domain were so far detected almost exclusively in bacteria and eukaryotes. This analysis indicates that caspase–hemoglobinase-fold proteases and their inactivated derivatives are widespread in diverse bacteria, particularly those with a complex development, such as *Streptomyces*, *Anabaena*, *Mesorhizobium*, and *Myxococcus*. The eukaryotic separin family was shown to be most closely related to the mainly prokaryotic HetF family. The phyletic patterns and evolutionary relationships between these proteins suggest that they probably were acquired by eukaryotes from bacteria during the primary, promitochondrial endosymbiosis. A similar scenario, supported by phylogenetic analysis, seems to apply to metacaspases and paracaspases, with the latter, perhaps, being acquired in an independent horizontal transfer to the eukaryotes. The acquisition of the caspase–hemoglobinase-fold domains by eukaryotes might have been critical in the evolution of important eukaryotic processes, such as mitosis and programmed cell death. Proteins 2002;46:355–367.**

Key words: caspase; apoptosis; mitosis; proteobacteria; cyanobacteria; cysteine-protease

INTRODUCTION

Caspases were first identified as thiol proteases involved in the regulation of inflammatory signaling and apoptosis in animals.[1–3] The caspase family was characterized by a proximal histidine and a distal cysteine in its active site.[4] The first X-ray structures of the caspases showed that these active site residues were positioned at the ends of strands that were embedded in a unique α/β-fold that did not resemble the folds found in other proteases.[5,6] Subsequent studies on the active site of eukaryotic vacuolar endopeptidases, also known as legumains or hemoglobinases, suggested the presence of a histidine–cysteine pair in a context similar to that in the caspases, suggesting that these proteases might have the same fold.[7] The same active site configuration was also observed in several diverse thiol proteases, including gingipains from *Porphyromonas gingivalis* and clostripains from *Clostridium* sp. (both pathogenic bacteria),[7] and more recently in the eukaryote-specific sister chromatid-separating proteases, separins.[8] Parallel studies using iterative sequence searches not only confirmed the relationship between caspases and legumains but also resulted in the detection of two additional families of caspase-related predicted proteases, the paracaspases and the metacaspases.[9–11] Taken together, these findings suggest that the class of caspase-related proteases has greater diversity than previously appreciated. The determination of the crystal structure of the *P. gingivalis* arginine gingipain helped in clarifying the conserved structural core of the common fold [hereinafter termed the caspase–hemoglobinase fold (CHF)] shared by these proteases.[12]

In functional terms, the caspases are the best studied of the CHF proteases because of their central role in cell death and inflammation. Caspases typically cleave substrates at DEXD motifs after the first aspartate.[3,4] The rare bacterial secreted proteases of this class, such as the gingipains and clostripains, have been chiefly studied in terms of their roles as virulence factors in bacterial pathogenesis.[7] Gingipains form a tight protein family;

[†] Correspondence to: L. Aravind, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD. E-mail: aravind@ncbi.nlm.nih.gov

Received 8 June 2001; Accepted 16 October 2001

they cleave polypeptides at R-X or K-X dipeptides and are accordingly differentiated into arginine and lysine gingipains.[12] The hemoglobinase family, in addition to the vacuolar endopeptidases, also includes transamidases that catalyze the addition of GPI anchors to proteins in a reversal of the proteolytic reaction.[13,14] The proteases of this family have been studied in the context of vacuolar protein degradation, antigen processing for presentation by MHC class II antigens, and assimilation of hemoglobin by blood parasites such as *Schistosoma*.[15–17] The characterized members of this family are asparagine-specific proteases.[16] The separins (separases) cleave the Scc1p subunit of the cohesin complex, allowing the separation of sister chromatids in anaphase.[18] Their target sequences typically contain the motif EXGR, with the cleavage occurring after the terminal arginine; in this respect, separins resemble R-gingipains.[18]

Of these protease families, hemoglobinases and separins are known to be highly conserved in all eukaryotes but so far have not been detected in prokaryotes. Caspases, paracaspases, and metacaspases have been previously shown to form a distinct group whose members are more closely related to one another than to the rest of the CHF proteases. Caspases are restricted in their distribution to the animal lineage, paracaspases have been detected in animals and *Dictyostelium*, and metacaspases have been found in plants, fungi, diverse early branching eukaryotes, and some bacteria, but not animals.[10,11] In contrast, clostripains and gingipains are limited to a few bacterial lineages, with no close homologs found elsewhere. This unusual phyletic distribution and the conservation pattern of the CHF proteases suggest a complex evolutionary history.

Here, we use the genome sequence data from diverse organisms and advanced sequence analysis methods, together with a cladistic approach, in an attempt to reconstruct some of these evolutionary events. We describe previously undetected CHF proteases, including a new family, and propose a coherent evolutionary classification for the proteins containing this fold.

RESULTS AND DISCUSSION

Detection of New CHF Proteases

An important problem associated with the CHF class of proteases is that, apart from the shared sequence signature of conserved catalytic histidine and cysteine residues, they show minimal sequence conservation between families and, therefore, are difficult to analyze with conventional sequence searches alone. Therefore, to extract proteins containing this fold from sequence databases as completely as possible, we applied a combination of iterative sequence profile searches, secondary structure prediction, and pattern and motif searches initiated with various seeds. Initial PSI-BLAST searches,[19] which were run to convergence with a profile inclusion *E*-value threshold of 0.01, were seeded with representatives of all previously identified CHF families, including hemoglobinases, caspases, paracaspases, metacaspases, separins, gingipains, and clostripains. The searches initiated with the hemoglobinase sequence, in addition to various eukaryotic members of this family, identified two previously undetected bacterial members from *Pseudomonas aeruginosa* and *Caulobacter crescentus* (PA4016 and CC2104, respectively). The searches initiated with the sequences of separins and gingipains did not detect any new CHF members other than orthologs or closely related paralogs from recently sequenced genomes. Searches with clostripain sequences detected, at convergence, multiple related predicted proteases from *Thermotoga maritima*, thereby extending the phyletic range of this family beyond the *Clostridia*. Searches started with the caspase, metacaspases, and paracaspase sequences recovered each other with statistically significant *E* values ($E < 0.001$ on first detection) and also identified a variety of previously uncharacterized proteins from the bacteria *Mesorhizobium loti*, *Streptomyces coelicolor*, *Myxococcus xanthus*, *Xylella fastidiosa*, and *Rhizobium* sp. (e.g., in a search started with the human paracaspase protease domain, these bacterial sequences were detected in iterations 3–5, with *E* values between 10^{-3} and 10^{-8}). Subsequent transitive iterations with these proteins recovered the hemoglobinase family members with *E* values ranging from 10^{-2} to 10^{-3}.

In addition, at convergence, some of these searches showed marginally significant hits ($E \sim 0.1$) to several uncharacterized proteins, such as Vng1314h from *Halobacterium* sp. and CC0601 from *C. crescentus*, which showed conservation not only of the histidine and cysteine catalytic pair associated with the preceding (predicted) strands but also of other N-terminal regions characteristic of the CHF. A search that began with the CC0601 protein of *C. crescentus* showed that these proteins belonged to a previously undetected family that, apart from uncharacterized proteins, included the HetF protein involved in heterocyst formation in the cyanobacterium *Nostoc*.[20] The majority of the members of this family (hereinafter the HetF family) come from diverse bacteria, but several were detected in eukaryotes and the archaeon *Halobacterium*. Most of the HetF family proteins retain the histidine–cysteine catalytic dyad typical of the CHF. The candidacy of these proteins to the CHF was further supported by the detection of members of this family, including Vng1314h and sll0638 (from *Synechocystis*), with *E* values ranging from 10^{-3} to 10^{-5}, in a search with a profile[21] that was constructed from a multiple alignment of all previously identified CHF proteins. A Gibbs sampling search[22,23] for conserved motifs in the entire set of the CHF proteins revealed the presence of three conserved motifs shared by the HetF family with the CHF proteins, with the probability of chance occurrence less than 10^{-8} for each motif. Secondary structure prediction for aligned proteins of the HetF family revealed a pattern of structural elements compatible with the CHF (as discussed later). A fold assignment for this family was further supported by the results of sequence–structure threading with the hybrid fold recognition method,[24] which detected 1PAU as the

structure best compatible with the HetF family sequences with moderate scores in the range of 10–15.

Taken together, these observations indicate that the HetF family proteins indeed contain CHF domains. The detection of new CHF-containing proteins allowed us to more precisely define the structural core of these domain proteins and to develop an evolutionary classification of this fold on the basis of interfamily- and intrafamily-specific sequence and structural features that, in evolutionary terms, are likely to represent shared derived characters (synapomorphies).

Conserved Structural Core of the CHF

To construct an optimal multiple alignment of the CHF proteases, we first aligned the individual families, including the newly detected ones, with the T_Coffee program;[25] this was followed by refinements on the basis of the PSI-BLAST search results. With these alignments as queries, the secondary structure was predicted for the individual families with the PHD program.[26,27] The alignments of the individual families were then combined, with the superposition anchored at the conserved motifs detected with the Gibbs sampling procedure and the predicted secondary structure elements, to generate a multiple alignment for the entire fold. The structural alignment between human caspases 1 and 3 and the arginine gingipain was generated with the FSSP database[28] and used as a further guide to refine the alignment between the secondary structure elements (Fig. 1).

The multiple alignment of the CHF protease sequences shows that conservation is centered around three prominent sequence motifs that correspond to an N-terminal β-strand, the central strand (strand 1) preceding the catalytic histidine, typically followed by a small residue, and a C-terminal strand (strand 4) preceding the catalytic cysteine (Figs. 1 and 2). These three strands, along with another, poorly conserved strand (strand 3), form a four-stranded parallel sheet at the core of the CHF domain, with the 2–1–3–4 topology from right to left (Fig. 2). The conserved core shared by all CHF proteins also contains three helices, one after strand 1, the second one after strand 2, and a very short one located between strands 3 and 4, thereby defining a simple α/β-fold.[5,6,12] Outside of the conserved core, which ends several amino acid residues to the C-terminus of the catalytic cysteine (motif III), no sequence conservation could be observed between most CHF protein families. The N-terminal strand of the CHF (motif I) often corresponds to the beginning of the polypeptide or occurs immediately after a distinct N-terminal domain. This defines the N-terminus of the minimal CHF domain. To verify the C-terminal boundary of this domain, distance matrix alignment (DALI) and vector alignment search tool (VAST) searches were run with the minimal conserved unit from caspases and R-gingipain. These searches showed that R-gingipain had a second copy of the CHF domain immediately N-terminal to the catalytic domain[12] (Figs. 1 and 2). This second repeat contained the same four-strand, three-helix core as the CHF unit identified by sequence comparisons, and despite the limited sequence conservation, a structure-based alignment with other CHF domains showed the presence of all sequence elements typical of this fold. An examination of the alignment of R-gingipains with K-gingipains showed that the latter also contained an N-terminal CHF domain. These observations strongly supported the aforementioned definition of the minimal CHF. The N-terminal CHF domain of R-gingipains retains the catalytic histidine and cysteine, but they are oriented differently (away from each other) from what is seen in the catalytic domains of gingipains and caspases.[12] Furthermore, the corresponding N-terminal CHF domain of K-gingipains lacks the catalytic residues, indicating that this domain is inactive. The N-terminal inactivated CHF domain of gingipain forms a covalently linked dimer with the C-terminal, catalytically active CHF domain in an arrangement that resembles caspase dimers.[5,6] This similarity suggests that dimeric interaction may have an ancient regulatory function in the CHF proteases.

Beyond the minimal common core, CHF domains show great diversity in terms of the conserved residues in helices, the lengths of the helices, and the inserts into the core (Fig. 1). The C-terminal extensions that tend to be conserved within CHF protease families are likely to form distinct structures that pack against the core sheet of the CHF domain and create an additional ligand-binding surface that might contribute to the specificity of these proteases. The CHF domain is a highly mobile domain that has combined, in the course of evolution, with a variety of other domains. Therefore, even proteins containing closely related CHF domains in some cases show major size differences and have distinct domain architectures (as discussed later); this makes domain architecture a poor indicator of deep evolutionary relationships between these proteins. We explored the distinctive features of CHF domains in detail to develop a classification of the CHF proteases presented next.

HetF: A Previously Undetected Family of CHF Proteases

The HetF family, a group of CHF-domain proteins that has not been described previously, is represented by multiple members in the bacteria *Synechocystis* sp. and *S. coelicolor* and the archaeon *Halobacterium* sp. and by a single member in the bacteria *C. crescentus* and *Nostoc punctiforme* (for which only a partial genome sequence is available). The three motifs typical of the CHF are strongly conserved in all prokaryotic HetF family proteins; however, many of them showed substitutions of the glycine in the position immediately after the catalytic histidine, which is otherwise nearly invariant in the CHF proteins (Fig. 1). Members of the HetF family were also detected in eukaryotes, namely, *Homo sapiens*, *Drosophila melanogaster*, and the early branching protist *Leishmania*, but despite the highly significant overall sequence similarity to the prokaryotic members, they all lack the conserved

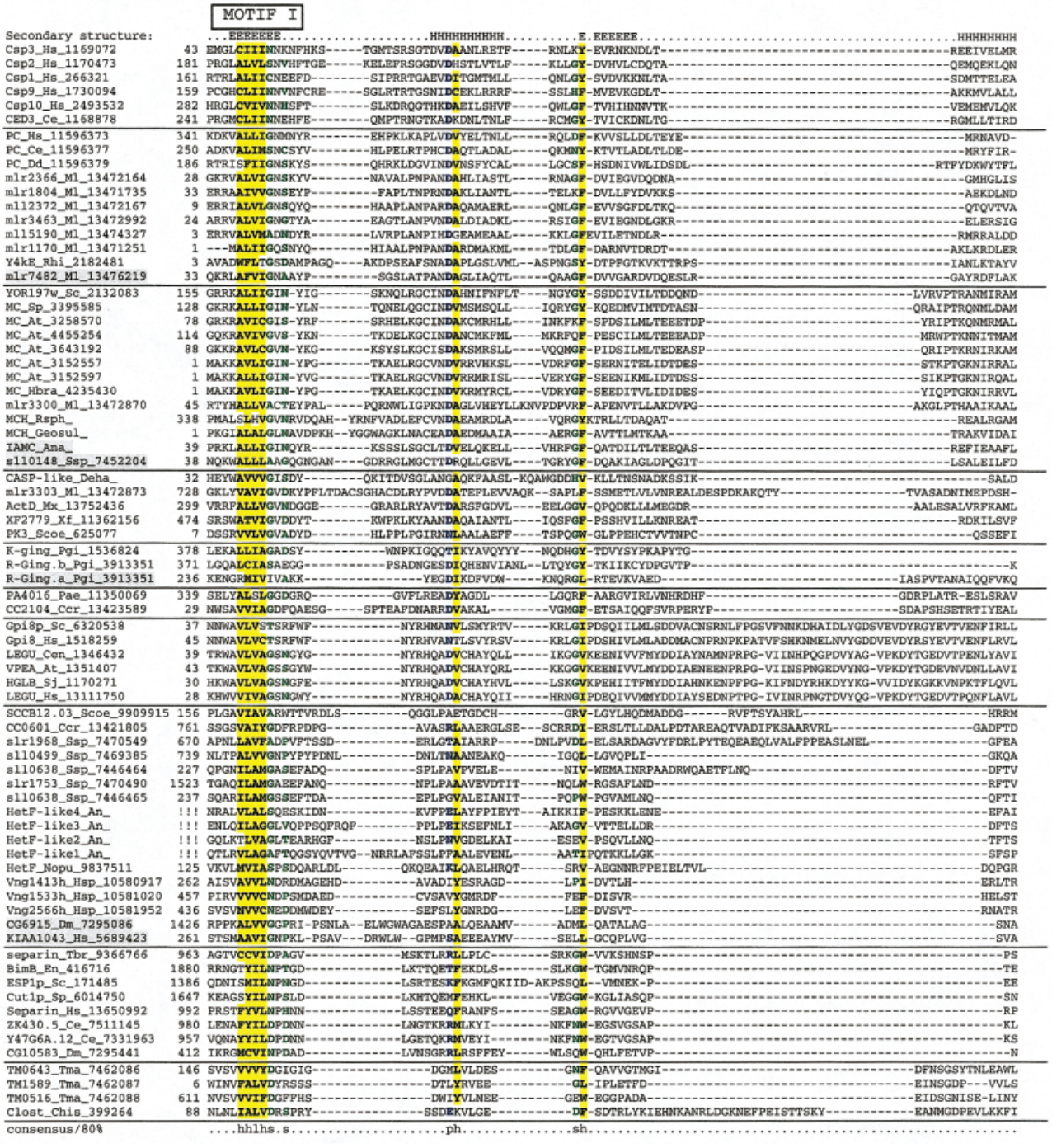

Fig. 1. Multiple alignment of a representative set of the CHF proteins. The multiple alignment was constructed as described in the text, and the secondary structure shown above the alignment was derived with the structures of caspase-3 (PDB: 1PAU) and R-gingipain (PDB: 1CVR). E indicates a β-strand, and H indicates an α-helix, with the upper case used to denote the common core likely to be present in the majority of the structures. The 80% consensus shown below the alignment was derived with the following amino acid classes: polar (p: KRHEDQNST), colored blue; hydrophobic (h: ALICVMYFW) and the aliphatic subset of these (l: ALIVMC), all shaded yellow; small (s: ACDGNPSTV), colored green, and the tiny subset of these (u: GAS), shaded green; and big (b: Q,E,R,K,Y,M,F,W,L,I), shaded gray. Catalytic residues are highlighted in reverse shading. The limits of the domains are indicated by the position numbers on each side of the alignment, with certain long inserts replaced by numbers within the alignment. The sequences are denoted by their gene names followed by the species abbreviations and GenBank identifiers. The names of the catalytically inactive domains are highlighted in gray. For sequences extracted from incomplete genomes, no GI numbers or domain boundaries are indicated. The different families of the CHF are separated by horizontal lines. From top to bottom, the families are caspases, paracaspases, metacaspases, generic PMC-related proteins, gingipains, bacterial hemoglobinase-like proteins, eukaryotic hemoglobinase–Gpi8-like proteins, the HetF family, separins, and clostripains. The species abbreviations are as follows: At, *Arabidopsis thaliana*; Cen, *Canavalia ensiformis*; Hs, *H. sapiens*; Dm, *D. melanogaster*; Ce, *Caenorhabditis elegans*; Sj, *Schistosoma japonicum*; Dd, *Dictyostelium discoideum*; Sc, *Saccharomyces cerevisiae*; Sp, *Schizosaccharomyces pombe*; En, *Emericella nidulans*; Hsp, *Halobacterium* sp.; An, *Anabaena* sp.; Cc, *C. crescentus*; Chis, *Clostridium histolyticum*; Dhal, *Dehalococcides ethenogenes*; Geosul, *Geosulfurococcus* sp.; Pa, *P. aeruginosa*; Rsph, *Rhodobacter spaehorides*; Scoe, *S. coelicolor*; Bpe, *B. pertussis*; Ml, *M. loti*; Mx, *M. xanthus*; Rhi, *Rhizobium* sp.; Pgi, *P. gingivalis*; and Nopu, *N. punctiforme*.

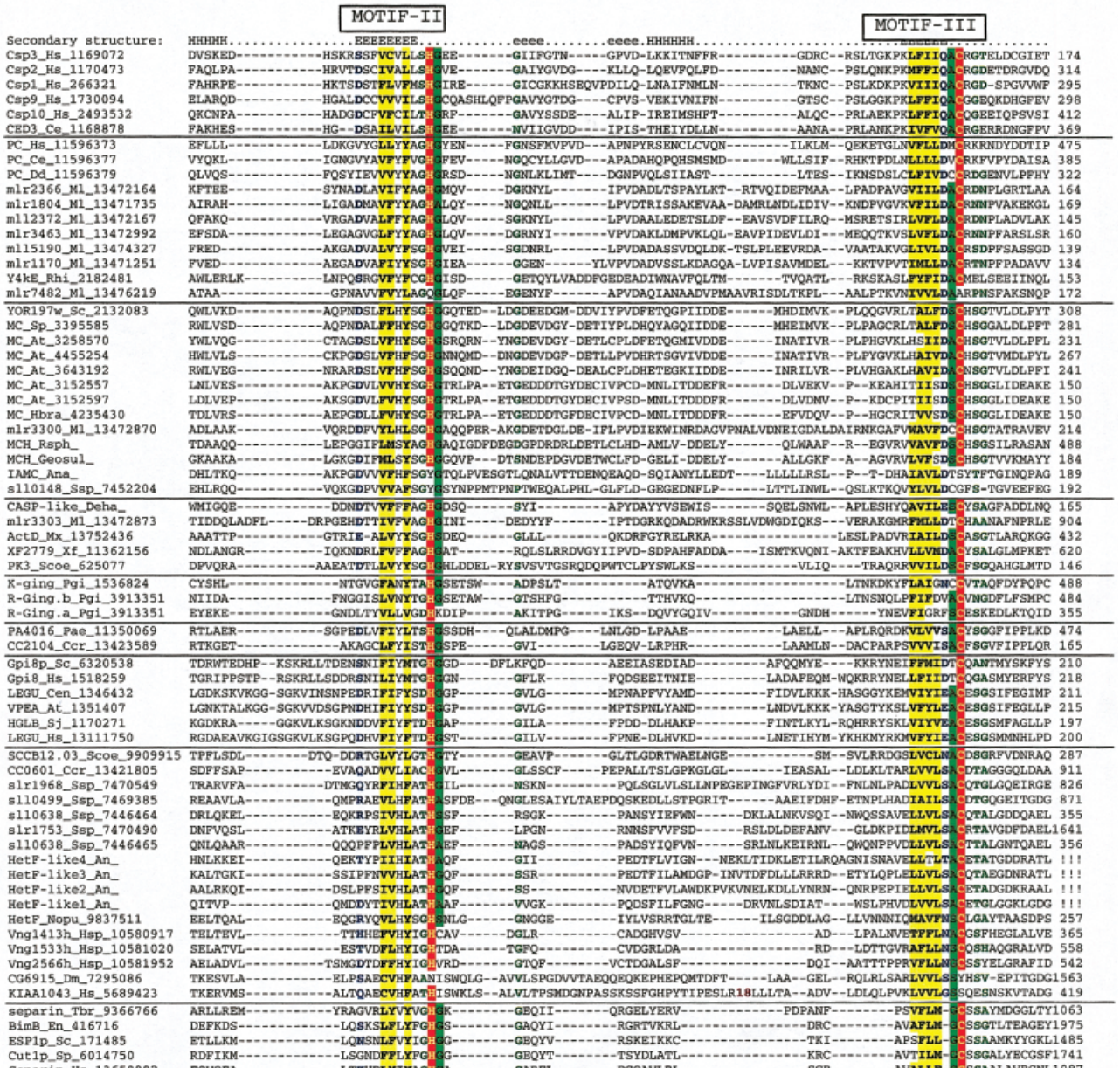

Figure 1. (Continued.)

histidine and/or cysteine (Fig. 1, Table I), indicating a secondary loss of the protease activity.

Database searches with a profile including all the HetF family members revealed significant similarity to eukaryotic separins (e.g., human separin was detected with an *E* value of 10^{-4} in the second iteration). These alignments showed the correct superposition of the catalytic C and H between the HetF and separin families and also revealed an additional region of conservation C-terminal to the CHF domain that was unique to these two families (Figs. 1 and 3). This C-terminal extension (hereinafter SepHet-C domain) appears to form a distinct folding unit that is analogous (but not homologous) to the specific C-terminal extensions present in other CHF families. The SepHet-C domain has a distinct pattern of conservation with a characteristic N-terminal motif, occurring between a (pre-

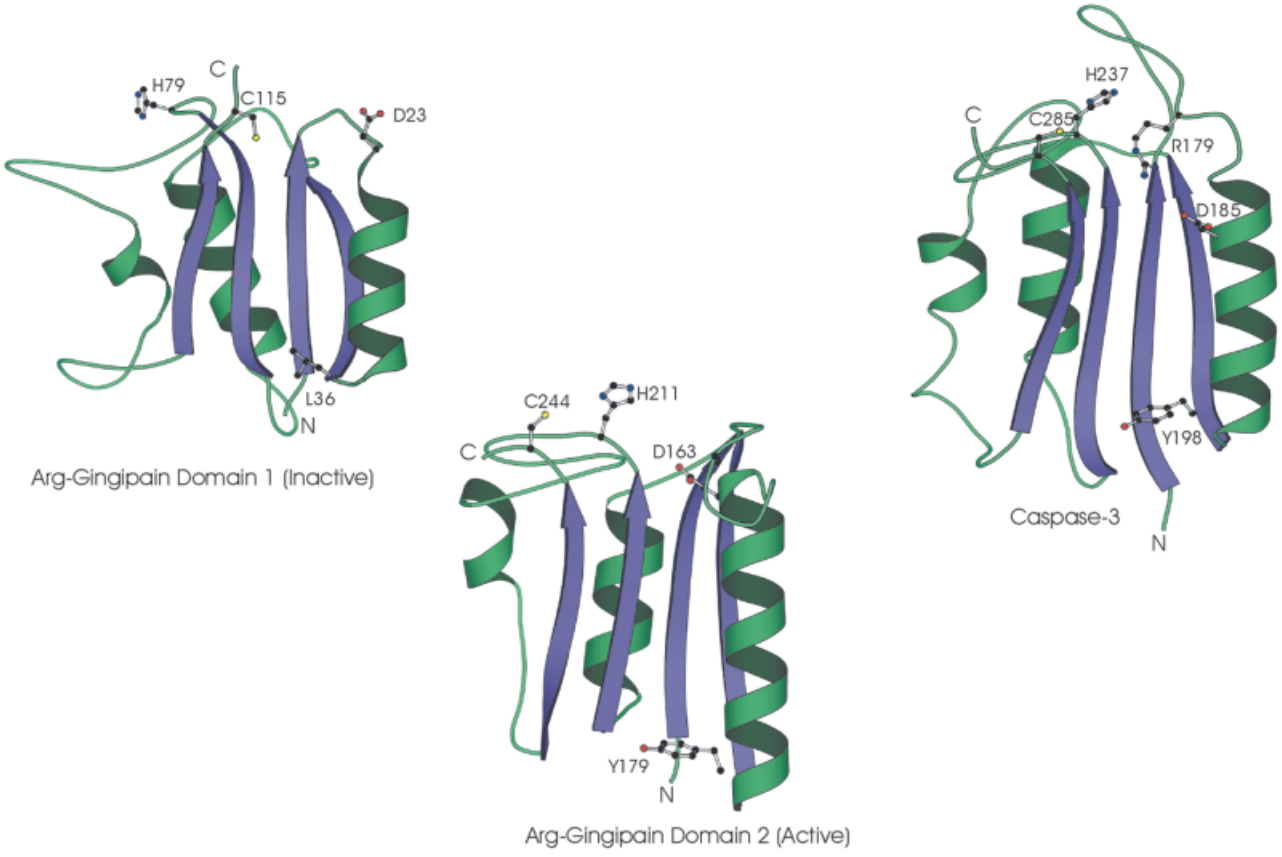

Fig. 2. Ribbon representation of the minimal core of the CHF. The minimal protease domains of caspase-3 (PDB: 1PAU) and the active and inactive copies from gingipain (PDB: 1CVR) are depicted. The catalytic residues and a selection of other equivalent conserved residues are depicted in the ball and stick format. In caspase-3, R179, which has a role in the interaction with the negatively charged substrate peptides, is indicated. Note the different orientation of the equivalents of the catalytic residues in the inactive copy of the domain in gingipain with respect to the two active copies. Also note the outward flip of the region C-terminal to strand-3 in the inactive copy.

```
Secondary Structure          .hHHHHHh....eEEEEe.....HHHHHHHHHHHHHHHHH.............HHHHHHHHHh.............eEEEEe
slr1968_Ssp_7470549      830 GLTRGFMYAGAARVVVSLWSVDDEATARLMTEFYRALLQRQMPPAAA    LQWAKQKLKEDPR    --FASPYFWAGF  899
sll0499_Ssp_7469385      874 GLSRSLIAAGVPSILVSLWSVDDASTEKLMTEFYQKWQQQGLGKAAA    LRQAMLQLKEE--    --YPEPYYWGAF  941
sll0638_Ssp_7446464      356 GFAGLALKAGVKSAVASFWNVDDAATLVLMTEFYRQLGQTSTKAEA-    LRQAQLKMLRGEL 25 TNFSAPYYWASF  451
slr1753_Ssp_7470490     1642 GFAGLAVKTGVKTAIGSLWYVSDEATFALMTSFYDRLQPAPIKAKA-    LQQAQLSLLRGEI 30 QDLSHPYFWSGF 1742
sll0638_Ssp_7446465      357 GFVGLALQSGVKSALGSLWRVSDLGTMALMSEFYWQLKAKSFKAQA-    LQQAQIRMINGEI 25 PELSHPYYWASF  452
SCCB12.03_Scoe_9909915   294 GFAELFLRKGAGGCIVSSGKVGDLEARAMARRLVREVAEHPRRPV--    VRTLRDFRVRALE 35 HYYGHPGTALRL  398
Vng1413h_Hsp_10580917    360 GLAL--VEGLSVAGAVTLRKVLDAQAAKVGTAFARLLLA-GFPIALA    LRLARRRIRMGKD 13 TDDRETGVAHVD  441
Vng1533h_Hsp_10581020    553 GRAL--VDRGSQGGIVTVSDIDNDAAVRVGHAIARLLND-GFSLLAG    LSVARNVTAFGNQ 13 SRSGMPLLIELA  634
Vng2566h_Hsp_10581952    537 GRAF--IDAGSQAGIITVRPIGNDAAMLIGQNIARFVNQ-GFSPRSA    LAIVQRYGRADDK 13 SKSGIPVVAHLT  618
HetF_Nopu_9837511        264 NLAESLVKRGIRSVLAMSERIPDEVALTLTQLFYRNLSQ-GYPVDLC    VSRVRQGLISAYG  3 MYWALPILYLQP  337
L5174.06_Lm_11065769    1349 GLLRGFFGSGVPCVIAGQWCTPDMKPMELFRYFYAQRCR-PLPTSPA  1 LSARTDTASGGSV 48 AMRYRPRVWAGY 1468
CC0601_Ccr_13421805      924 GLARGFIYAGASDVLATQWKVDSASSAAQMHAFFENAGA-ASQPLAK  1 LASSQRSLYSNPE    --TGHPFYWAAF  993
KIAA1043_Hs_5689423      422 ALTRAFLAAGAQCVLVSLWPVPVAASKMFIHAFYSSLLN-GLKASAA    LGEAMKVVQSSKA    --FSHPSNWAGF  490
CG6915_Dm_7295086       1566 QLAGGWLLAGAGVVLISLWPVPETAAKILLRAFYSALLQ-GARAARA    LAEAMQTVQHTKH    --FAHPANWAGF 1634
separin_Tbr_9366766     1067 GMPYAFLHAGAPLFVGCLWHVTDGEIDRLTKRLLSFV---SYGGGSG 14 LRLARKSCKLPYL  5 VLYGMNLPLGGT 1154
BimB_En_416716          1979 GTPMNYLQAGSPALVATLWDVTDKDIDRFAKATFEHW---GLIGNGH 20 RGACVLKYLNGAA  1 VVYGVPGVFLH- 2067
ESP1p_Sc_171485         1489 GTIYTYLLGGCPMVLGNLWDVTDKDIDKFSEELFEKM---GFRCNTD 16 RGVCHLRYLNGAA  1 VIYGLPIKFVS- 1573
Cut1p_Sp_6014750        1745 GTPLDYLSAGCPTLVANLWDVTDKDIDRFSLKMLESW---GLFENKA 15 RSCCHLRYLNGAA  1 VIYGIPAYIIP- 1828
Separin_Hs_13650992     1091 GIVLKYIMAGCPLFLGNLWDVTDRDIDRYTEALLQGWL--GAGPGAP  8 RQAPRLKYLIGAA  1 IAYGLPVSLR-- 1167
Y47G6A.12_Ce_7331963    1056 TAILDYAMAKCPLIVGCLWTVTDGEIDRFLIRMIDDCF--EDSKSLT 15 RSKARLKYLTGAA  1 VMYGLPVVAKQT 1141
CG10583_Dm_7295441       513 GAHDYYHGALCPSIVGTLMPALDGNMDTVSVTILSRWLAPGDNKVIP  4 DRVPWLKNGIIKG 41 VCRGLPAWNLAV  629
consensus/90%                s.....h..Gs...lss.b.lss.....h...hhp.h...s.....    ...s......s..    ...s.s..h...
```

Fig. 3. Multiple alignment of the SH. The alignment was constructed and colored as in Figure 1. The HetF family protein sequences are shown at the top of the figure, and separin sequences are shown at the bottom. The species abbreviations are listed in Figure 1.

dicted) strand and a helix, which contains a conserved aromatic residue followed by an aspartate three residues downstream (Fig. 3). The conservation of the SepHet-C domain in the HetF and separin families, together with the overall sequence similarity and the presence of shorter loops and helices compared with other CHF proteins (Fig.

TABLE I. Phyletic Distribution of Families of the CHF†

Family	Bacteria	Archaea	Early branching eukaryotes	Crown group eukaryotes
		Caspasoid class		
		PMC subclass		
Caspases	—	—	—	4–15 copies per genome; seen thus far only in animals
Paracaspases	*Mesorhizobium loti* (7), *Rhizobium* (1)	—	—	*Dictyostelium, Caenorhabditis, Homo* (1 copy each)
Metacaspases	*Anabaena* (2), *Synechocystis* (1), *M. loti* (1), *Geococcus* (at least 1), *Rhodosphaera* (1)	—	*Trypanosoma, Plasmodium* (at least 1 copy)	Fungi (at least 1 copy per genome), plants: *Arabidopsis* (10 copies)
Generic PMC members	*Myxococcus xanthus* (at least 1), *Anabaena* (5), *M. loti* (1), *Xylella fastidiosa* (1), *Streptomyces coelicolor* (1), *Bordatella pertussis* (1), *Dehalococcoides ethenogenes* (1), *Pseudomonas syringae* (1)	—	—	—
Gingipains	*Porphyromonas gingivalis* (2 copies)			
		Hemoglobinase-like subclass		
Eukaryotic hemoglobinases	—	—	*Plasmodium, Trypanosoma* (1–2 copies at least)	All crown group eukaryotes show at least 1 transamidase and 1 vacuolar protease homolog.
Bacterial hemoglobinase-like proteins	*Pseudomonas aeruginosa* (1), *Caulobacter crescentus* (1)	—	—	—
		Separinoid class		
Separins	—	—	*Plasmodium* (1), *Trypanosoma* (1)	1 copy in all crown group eukaryotes; 2 in *C. elegans*
HetF	*Synechocystis* (5), *Nostoc* (at least 1), *S. coelicolor* (2), *C. cresentus* (1), *Anabaena* (19)	*Halobacterium* (3)	*Leishmania* (1)	*Homo sapiens* (1), *Drosophila melanogaster* (1)
		Unassociated members		
Clostripains	*Clostridium* (1), *Thermotoga* (3)	—	—	—

†The number of detected representatives of each family is indicated in parentheses after the species name.

1), suggests that the two families are sister groups within the CHF class of proteases.

HetF is a positive regulator of the development of nitrogen-fixing heterocysts in the filamentous cyanobacterium *N. punctiforme*.[20] This pathway also includes the self-degrading serine protease HetR,[20] and it seems likely that HetF initiates a proteolytic regulatory step, either in conjunction with or upstream of HetR, to trigger the heterocyst-specific gene expression. The presence of HetF homologs in a variety of bacteria suggests that these proteases function in proteolytic signaling pathways in diverse contexts. The presence of a single inactive HetF-like protein in several eukaryotic lineages points to an ancient, conserved function. Inactive CHF proteases form noncovalent or covalent dimers with their active homologs (e.g., the covalent dimer in gingipain discussed previously), in which the inactive copy typically acts as a negative regulator.[11,29,30] Given the ubiquitous presence, in eukaryotes, of the typically single-copy separins, which appear to be related to the HetF family, it seems plausible that the inactive HetF-like proteins are dominant-negative regulators of the separins, with which they could form heterodimers. In contrast to the ubiquitous separins, HetF family proteins show sporadic distribution in eukaryotes,

which suggests that the proposed mechanism of negative regulation has been lost or displaced in other lineages. Two of the *Synechocystis* HetF-like proteins and the human and *Drosophila* ones contain long, N-terminal extensions with multiple copies of tetratricopeptide repeats (TPRs; Fig. 4), indicating that they might form scaffolds of multiprotein complexes.

New Members of the Paracaspase–Metacaspase–Caspase (PMC) Subclass of CHF Proteases

In addition to the previously described bacterial members of this subclass of the CHF from *Streptomyces*, *Anabaena*, *Xylella*, *Rhizobium*, *Bordetella pertussis*, and *Geosulfurococcus*,[10,11] we detected a particularly diverse set of CHF proteins in the recently sequenced genome of *M. loti* and the developmentally complex bacterium *M. xanthus* (for which a partial genome sequence is available). The previously detected bacterial CHF proteins, such as Y4kE from *Rhizobium* and PK3 from *S. coelicolor*, are roughly equidistant from the metacaspase and paracaspase families in terms of sequence conservation patterns. A similar general relationship of the metacaspases and paracaspases was observed for the newly detected ActD protein from *M. xanthus*, for XF2779 from *X. fastidiosa*, and for one of the *M. loti* CHF proteins, mlr3303. The *M. loti* protein mlr3300 and related proteins from *Anabaena*, *Geosulfurococcus*, and *Rhososphaera* showed a closer relationship with metacaspases in terms of sequence similarity and shared sequence patterns in the second and third strands and the region around the catalytic cysteine (Fig. 1). In contrast, the other seven CHF proteins from *M. loti* were obviously related to paracaspases, especially with respect to specific sequence features of the first strand, first helix, and the intervening regions downstream of strands 3 and 4 (Fig. 1). *Anabaena* and *Synechocystis* encode metacaspase-like proteins that lack the catalytic residues (IAMC and sll0148, respectively), whereas *M. loti* encodes a similarly inactivated paracaspase derivative (mlr7482; Fig. 1). These observations point to the independent emergence of inactive variants of CHF proteases with a potential regulatory role on several occasions during evolution.[11]

Many of the newly detected bacterial members of this subclass of CHF proteases showed complex architectures with fusion to diverse repetitive domains (Fig. 4). The paracaspase-like protein mlr1804 contains C-terminal Sel-1 repeats, and mll5190 and mlr1170 contain TPRs, also at the C-terminus, whereas mlr3303 has N-terminal WD40 (WD) repeats (Fig. 4). Two of the *Anabaena* metacaspase-like domains are fused to apoptotic ATPase (AP-ATPase) domains, followed by C-terminal WD repeats, whereas the previously described member of this family from *S. coelicolor*, PK3, is fused to a protein kinase.[9,11] These domain architectures suggest that bacterial CHF proteases function within signaling complexes, in which they associate with other proteins through repetitive, superstructure-forming domains. Several of the bacteria that encode caspase-like proteases (*S. coelicolor*, *M. loti*, *Synechocystis*, *Anabaena*, and *M. xanthus*) also encode AP-ATPases[9] or NACHT-NTPases[31] and eukaryote-type protein kinases of the PKN2 family,[32] all typical components of eukaryotic signaling systems and, in particular, the apoptotic system. The fusion of the metacaspase-like domain with AP-ATPase in *Anabaena* is an especially telltale observation, which strongly suggests that some of the bacterial homologs of eukaryotic apoptotic proteins functionally interact in bacterial signaling systems. Five of the nine PMC-subclass proteins in *M. loti* and the sole member in *X. fastidiosa* contain predicted signal peptides (sp; Fig. 4). This suggests a role for these secreted proteases in interactions of the bacteria with their plant hosts.

The only bacterial member of the PMC subclass for which some direct functional information is available is the ActD protein encoded in the act operon of *M. xanthus*, which is involved in regulation of the sporulation morphogen CsgA.[33] The act operon appears to encode two distinct regulatory systems, with ActA (a protein that consists of a receiver domain and a GGDEF nucleotide cyclase domain) and ActB (a NtrC-like, DNA-binding transcription factor) forming one pathway and ActC and ActD forming the other.[33] The latter system controls the timing of CsgA production. The ActC protein consists of a carbohydrate dehydratase domain fused to an NH_2–acetyltransferase domain and probably is involved in the biosynthesis of an uncharacterized oligosaccharide metabolite that regulates the expression of the csgA operon. The predicted protease activity of ActD could be involved in regulatory processing of ActC or associated proteins. This function is consistent with the fusion of the protease domain of the *X. fastidiosa* ActD homolog, XF2779, with an oligosaccharide deacetylase domain, which indicates that a similar regulatory mechanism could operate in this bacterium.

CC2104 and PA4016 and Their Relationship With the Hemoglobinase Family

The hemoglobinase family has so far been restricted in its distribution to eukaryotes, but in iterative searches with members of this family, two bacterial proteins, CC2104 from *C. crescentus* and PA4016 from *P. aeruginosa*, were detected. These proteins appear to form a small, bacteria-specific family. Both of them contain a predicted signal peptide, an active CHF domain, and a C-terminal extension, which consists of 60–70 residues and is specifically shared with the hemoglobinase family. This extension contains two sequence signatures, TAA and CXD (data not shown), and is likely to form a distinct structural unit that could contribute to the substrate specificity of the protease. However, these bacterial proteins differ from the eukaryotic hemoglobinases in terms of the conservation pattern in strand 2 and the presence of a large insert between strand 2 and helix 3 (Fig. 1). In *P. aeruginosa*, this predicted secreted protease potentially could have a role in pathogenesis.

Classification and Evolutionary History of the CHF

The CHF proteins were classified with a combination of similarity-based clustering, symmetry of recovery in PSI-BLAST searches, phylogenetic tree analysis for individual families, and cladistic analysis based on shared derived characters. With these approaches, two major classes of CH domains were identified; they were provisionally designated the caspasoids and the separinoids (Table I, Fig. 4). The caspasoid class consists of the PMC subclass and the hemoglobinase-like subclass and is unified by several synapomorphic features, such as the presence of a conserved aspartate at the N-terminus of helix 1 and a well-developed strand 2. The hemoglobinase-like subclass differs from the PMC subclass by the presence of a unique C-terminal extension (as discussed previously) and is further subdivided into the eukaryotic hemoglobinases (including the vacuolar proteases and GPI8) and the bacterial hemoglobinase-like proteins (CC2104 and PA4016). The PMC subclass is currently the largest assemblage within the CHF and consists of at least four distinct families, namely, caspases, metacaspases, paracaspases, and gingipains. Several bacterial members of this subclass are generically related to both metacaspases and paracaspases and might be most closely related to the common ancestor of these families (Figs. 4 and 5). The gingipains are distinct members of this subclass, but their remarkably restricted phyletic distribution (so far detected only in *Porphyromonas*) suggests that they have recently diverged from bacterial PMC proteins, probably in response to selection due to the host defenses.

The separinoid class includes two families, namely, the HetF family and the separin family, which are unified by the presence of the shared C-terminal SepHet-C module (SH; Fig. 3) and a divergent, shortened helix 1 and strand 2. The separin family shows an unusual deviation from the ancestral state of the CHF, having lost the otherwise conserved polar residue two positions upstream of the catalytic cysteine, and is clearly distinguished from the HetF-like proteins on the basis of this synapomorphy and overall sequence conservation (Fig. 1).

The clostripains are another group of secreted bacterial proteases from *Clostridium* and *T. maritima*. Analogous to the gingipains, they contain three previously unnoticed N-terminal immunoglobulin domains (IG; Fig. 4), which might mediate the attachment of these proteins to extracellular substrates. However, clostripains do not have any specific sequence features that would group them with the gingipains or any other proteins of the two major CHF classes. Nevertheless, their very limited phyletic distribution suggests that clostripains probably are highly divergent derivatives of one of the two classes (Fig. 4).

This classification of CHF proteins (Fig. 4), combined with their phyletic distribution patterns (Table I), provides for a reconstruction of their evolutionary history. This reconstruction, although being necessarily speculative, is the most conservative scenario compatible with the currently available data. The CHF proteins are widespread in both eukaryotes and bacteria, but, except for three closely related HetF-like proteins in *Halobacterium*, they are undetectable in the proteomes of diverse archaea. This suggests that the CHF first emerged either in the eukaryotic bacterial lineage, with a subsequent expansion of their phyletic horizon via lateral gene transfer. The alternative scenario would postulate that the CHF was already present in the common ancestor of all three major superkingdoms of life but was lost in most archaeal lineages. However, the representatives of the CHF present, so far, in a single archaeon, *Halobacterium*, do not appear to be relics of an ancestral CHF protein. These lack the diversity seen among the CHF proteins in either eukaryotes or bacteria and are specifically related to a single lineage (HetF) that is widespread in the bacteria (Fig. 4), which suggests acquisition of these genes by *Halobacterium* via horizontal transfer from a bacterial source.

In the separinoid class, both eukaryotic groups of proteins, the separins and the HetF homologs, are highly derived with respect to the apparent primitive state seen in the bacterial homologs (Fig. 1 and as discussed previously). Furthermore, they are present in a single copy, with the exception of the separin duplication in the nematode, as opposed to the considerable diversity seen in the bacterial HetF proteins (Fig. 4). A substantial set of eukaryotic proteins, unlike those involved in core processes, such as replication, transcription, and translation,[34,35] show phylogenetic affinities to bacteria instead of archaeal proteins.[36] It has been suggested that genes coding for many, if not most, of these proteins were acquired by eukaryotes from the promitochondrion, which was derived from an α-proteobacterial endosymbiont associated with the common ancestor of practically all known eukaryotes.[37,38] It has also been proposed that phagocytosis of bacteria by ancestral unicellular eukaryotes might have served as an additional conduit for the lateral acquisition of bacterial genes by the eukaryotic lineage.[39] These observations, taken together with the phyletic pattern discussed previously, suggest that the eukaryotic members of the separinoid class were probably derived from prokaryotic ones via horizontal transfer at an early stage of eukaryotic evolution. This was followed by extensive sequence divergence, perhaps triggered by colonization of new functional niches, leading to separins and the inactivated HetF-like proteins. The apparent conservation throughout eukaryotic evolution and the detection of a separinoid class protein in α-proteobacteria (*C. crescentus*) seem to be compatible with their entry into the eukaryotes occurring in the course of the primary endosymbiosis that gave rise to the mitochondrion. The available data do not allow us to infer the primary function of these proteases when first acquired by the eukaryotes. Given the critical role of separins in chromosomal organization and mitosis in diverse eukaryotes, it seems likely that they were recruited for this function shortly after the promitochondrial symbiosis.

The caspasoid class is also present in eukaryotes and bacteria but missing in diverse archaeal proteomes sampled to date. Phylogenetic tree construction for the PMC sub-

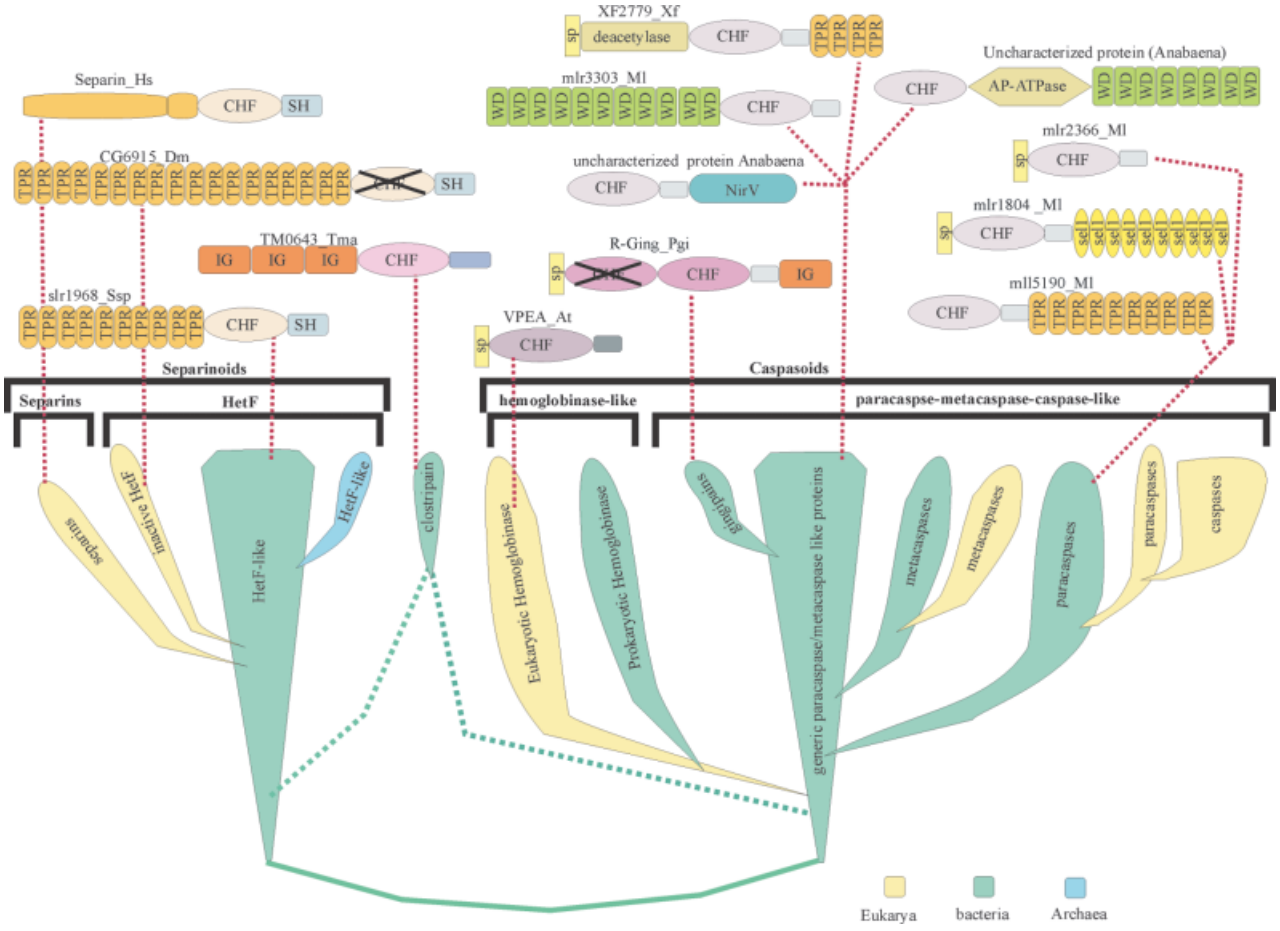

Figure 4

Figure 5

class with the neighbor-joining and maximum-likelihood methods suggested that all the eukaryotic members of this class were derived from within bacterial families. The eukaryotic paracaspase–caspase and metacaspase branches appeared to strongly group (bootstrap values > 70%) with distinct bacterial paracaspases or metacaspases, respectively, whereas several additional generic bacterial PMC subclass members remained outside these clusters (Fig. 5). Trees that place the eukaryotic PMC subclass members together, to the exclusion of the bacteria, showed significantly lower likelihood values[40] than those that placed the eukaryotic members within the bacteria (Fig. 5). Therefore, it appears most likely that metacaspases, paracaspases, and generic members of the PMC subclass had diversified in bacteria via a series of ancient duplications (Fig. 5). Given the conservation of metacaspases in all eukaryotes other than animals, including early branching lineages, and their presence in α-proteobacteria (Table I), they were probably acquired by eukaryotes from the promitochondrial endosymbiont. It seems likely that concomitant acquisition of other components of bacterial signaling systems that might be functionally linked to metacaspases and paracaspases already in bacteria, namely, AP-ATPases and NACHT-NTPases, had a major role in the evolution of the eukaryotic apoptosis system.[11]

Among eukaryotes, paracaspases so far have been found only in animals and slime molds.[10] The evidence for the differentiation of the paracaspases and metacaspases in bacteria, along with the presence of the former in a limited set of eukaryotes, implies that they might have been acquired by eukaryotes via a second horizontal transfer into the common unicellular ancestor of the *Dictyostelium*-animal lineage. This route of horizontal transfer is not unlikely because the precursor of the *Dictyostelium*-animal lineage probably was an amoeboid organism that phagocytosed bacteria.[36,41] However, the alternative, namely, that the paracaspases were also derived as a part of the massive gene transfer accompanying the initial promitochondrial endosymbiosis and have merely not been found in early branching eukaryotic lineages studied so far, cannot be ruled out with these data. The caspase family is restricted to animals and has a closer relationship with the paracaspase family, as indicated by several shared sequence features.[10] The diversity of the caspase family increased relatively late in animal evolution, particularly in the coelomates; therefore, this branch of the PMC subclass appears to be a distinct, animal-specific development.

The hemoglobinase-like subclass is the only major group of CHF proteins that shows a predominantly eukaryotic distribution. Hemoglobinase sequences are highly conserved throughout the eukaryotic domain. In contrast, the bacterial family of hemoglobinases has only two members that are closely related to each other, but they are distinct from the eukaryotic family. Horizontal transfer may have contributed to the evolution of this group also, but the direction, in this case, is currently impossible to ascertain. Given the limited sequence diversity of hemoglobinases and their clear relationship with the caspase-like proteins, it appears possible that they are ancient, highly derived forms that originally emerged from within the PMC subclass (Fig. 4).

CONCLUSIONS

The detailed sequence and structural analysis of the CHF proteins described here resulted in the delineation of the evolutionarily mobile structural core of the protease domain and the identification of many new members, including the previously undetected HetF family. Most CHF proteins are predicted to be active proteases, but multiple, independent inactivations were observed in almost all lineages within this fold. Taken together with the tendency to form intramolecular or intermolecular dimers, this suggests an important regulatory role for the inactive forms. A classification of the CHF was developed to reflect the inferred evolutionary relationships between the constituent protein families. The CHF proteins are so far almost completely limited to bacteria and eukaryotes in their phyletic distribution. They appear to have widely propagated in diverse bacteria, particularly those that undergo complex development, such as *Streptomyces*, *Anabaena*, *Mesorhizobium*, and *Myxococcus*. The evolutionary relationships and phyletic patterns of the CHF proteins suggest that they have been acquired by eukaryotes from bacteria via horizontal gene transfer. The principal source for this acquisition appears to be the promitochondrial endosymbiosis, but additional, subsequent horizontal transfers between bacteria and early, unicellular eukaryotes also appear possible. In eukaryotes, these CHF

Fig. 4. Inferred evolutionary scenario for the CHF proteases and domain architectures of the newly detected proteins containing the CHF domain. A hypothetical reconstruction of the evolution of the CHF was derived with a combination of a cladistic approach, sequence-similarity-based clustering, and the data on retrieval of different families in PSI-BLAST searches as described in the text. The dotted blue lines that connect clostripains with both the separinoid and caspasoid lineages reflect the uncertainty regarding the derivation of this group (see the text). The areas of the shapes indicating the lineages are roughly proportional to their diversity in extant organisms. The dotted red lines connect distinct branches of the tree with schematic depictions of the architectures of the newly detected CHF-domain-containing proteins of the corresponding groups. Only the globular domains are shown, roughly to scale. The different colors of the CHF represent different forms of this domain, and the gray rectangles shown in its C-terminus represent the conserved extensions shared by each group of CHF proteins. Inactivated forms of the CHF are indicated by black crosses. For uncharacterized proteins, systematic gene names are indicated wherever available. The species abbreviations are as follows: At, *Arabidopsis thaliana*; Dm, *D. melanogaster*; Hs, *H. sapiens*; Ml, *M. loti*; and Ssp, *Synechocystis* sp.

Fig. 5. Maximum-likelihood phylogenetic tree for the PMC subclass. An alignment of the PMC subclass (see Fig.1) was used to construct a distance matrix with the PROTDIST program of the PHYLIP package with the George–Hunter–Baker categories model. This matrix was used to construct a minimum evolution tree with the FITCH program, and this tree was subject to local rearrangements to find the maximum-likelihood tree with the PROTML program of the MOLPHY package. The Rell-BP bootstrap for each node was calculated on 10,000 resamplings of the data set. The eukaryotic branches are colored blue, whereas the bacterial branches are colored red, and the protein names follow exactly the same convention followed in Figure 1.

proteases assumed critical roles in essential eukaryotic processes, including chromosomal separation in mitosis and programmed cell death. The extreme sequence divergence seen in the CHF makes it possible that some members, especially inactive ones, have eluded the current detection methods. However, given the depth of the searches presented here, it appears likely that these potential undetected members are either highly derived or limited in their phyletic distribution.

MATERIALS AND METHODS

The Nonredundant Protein Sequence database, the Expressed Sequence Tags database (National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD), and the individual protein sequence databases of completely and partially sequenced genomes accessible at http://www.ncbi.nlm.nih.gov/Microb_blast/unfinishedgenome.html were searched with the gapped version of the BLAST programs[19,42] (BLASTPGP for proteins and TBLASTNGP for translating searches of nucleotide databases). Sequence profile searches were performed with the PSI-BLAST program; profiles were saved with the –C option and retrieved with the –R option.[19,21] Multiple alignments of amino acid sequences were generated with a combination of PSI-BLAST, T_Coffee,[25] Gibbs sampling,[22,23] and secondary structure predictions produced with the PHD program,[26,27] with multiple alignments of individual protein families used as queries. Phylogenetic analysis was carried out with the neighbor-joining algorithm, with subsequent local rearrangements with the maximum-likelihood algorithm. The likelihood of alternative positions for selected clades was assessed with the Kishino–Hasegawa test.[43] The packages used for these phylogenetic analysis were PHYLIP, PAUP, and MOLPHY.[40,44,45] Sequence–structure threading was carried with the combined fold prediction algorithm.[24] Signal peptides in protein sequences were predicted with the SignalP program.[46] The three-dimensional structure visualization, alignment, and modeling was carried out with the SWISS PDB Viewer program.[47] Ribbon diagrams were generated with Molscript.[48]

NOTE ADDED IN PROOF

Recent completion of sequencing of the genome of cyanobacterium Anabaena (Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A, Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S. Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium anabaena sp. strain PCC 7120. DNA Res 2001;8:205–213) allowed us to update the counts for the CHF-fold proteins in this organism. A dramatic expansion of the HetF family was observed, with a total of 19 distinct members encoded in the complete genome. Three of these proteins show a previously undetected fusion to an uncharacterized C-terminal domain, which is also fused to cNMP cyclases and protein kinases from a variety of bacteria. These observations indicate that at least some predicted proteases of the HetF family are involved in signal transduction.

REFERENCES

1. Nicholson DW, Thornberry NA. Caspases: killer proteases. Trends Biochem Sci 1997;22:299–306.
2. Earnshaw WC, Martins LM, Kaufmann SH. Mammalian caspases: structure, activation, substrates, and functions during apoptosis. Annu Rev Biochem 1999;68:383–424.
3. Salvesen GS, Dixit VM. Caspases: intracellular signaling by proteolysis. Cell 1997;91:443–446.
4. Stennicke HR, Salvesen GS. Catalytic properties of the caspases. Cell Death Differ 1999;6:1054–1059.
5. Wilson KP, Black JA, Thomson JA, Kim EE, Griffith JP, Navia MA, Murcko MA, Chambers SP, Aldape RA, Raybuck SA, Livingston DJ. Structure and mechanism of interleukin-1 beta converting enzyme. Nature 1994;370:270–275.
6. Walker NP, Talanian RV, Brady KD, Dang LC, Bump NJ, Ferenz CR, Franklin S, Ghayur T, Hackett MC, Hammill LD, Xiong L., Möller A. Crystal structure of the cysteine protease interleukin-1 beta-converting enzyme: a (p20/p10)2 homodimer. Cell 1994;78:343–352.
7. Chen JM, Rawlings ND, Stevens RA, Barrett AJ. Identification of the active site of legumain links it to caspases, clostripain and gingipains in a new clan of cysteine endopeptidases. FEBS Lett 1998;441:361–365.
8. Uhlmann F, Wernic D, Poupart MA, Koonin EV, Nasmyth K. Cleavage of cohesin by the CD clan protease separin triggers anaphase in yeast. Cell 2000;103:375–386.
9. Aravind L, Dixit VM, Koonin EV. The domains of death: evolution of the apoptosis machinery. Trends Biochem Sci 1999;24:47–53.
10. Uren GA, O'Rourke K, Aravind L, Pisabarro TM, Seshagiri S, Koonin VE, Dixit MV. Identification of paracaspases and metacaspases: two ancient families of caspase-like proteins, one of which plays a key role in MALT lymphoma. Mol Cell 2000;6:961–967.
11. Aravind L, Dixit VM, Koonin EV. Apoptotic molecular machinery: vastly increased complexity in vertebrates revealed by genome comparisons. Science 2001;291:1279–1284.
12. Eichinger A, Beisel HG, Jacob U, Huber R, Medrano FJ, Banbula A, Potempa J, Travis J, Bode W. Crystal structure of gingipain R: an Arg-specific bacterial cysteine proteinase with a caspase-like fold. EMBO J 1999;18:5453–5462.
13. Ishii S. Legumain: asparaginyl endopeptidase. Methods Enzymol 1994;244:604–615.
14. Meyer U, Benghezal M, Imhof I, Conzelmann A. Active site determination of Gpi8p, a caspase-related enzyme required for glycosylphosphatidylinositol anchor addition to proteins. Biochemistry 2000;39:3461–3471.
15. Manoury B, Hewitt EW, Morrice N, Dando PM, Barrett AJ, Watts C. An asparaginyl endopeptidase processes a microbial antigen for class II MHC presentation. Nature 1998;396:695–699.
16. Chen JM, Fortunato M, Barrett AJ. Activation of human prolegumain by cleavage at a C-terminal asparagine residue. Biochem J 2000;352:327–334.
17. Klinkert MQ, Felleisen R, Link G, Ruppel A, Beck E. Primary structures of Sm31/32 diagnostic proteins of Schistosoma mansoni and their identification as proteases. Mol Biochem Parasitol 1989;33:113–122.
18. Uhlmann F, Lottspeich F, Nasmyth K. Sister-chromatid separation at anaphase onset is promoted by cleavage of the cohesin subunit Scc1. Nature 1999;400:37–42.
19. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
20. Wong FC, Meeks JC. The hetF gene product is essential to heterocyst differentiation and affects HetR function in the cyanobacterium Nostoc punctiforme. J Bacteriol 2001;183:2654–2661.
21. Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF. IMPALA: matching a protein sequence against a collection of

PSI-BLAST-constructed position-specific score matrices. Bioinformatics 1999;15:1000–1011.
22. Schuler GD, Altschul SF, Lipman DJ. A workbench for multiple alignment construction and analysis. Proteins 1991;9:180–190.
23. Neuwald AF, Liu JS, Lipman DJ, Lawrence CE. Extracting protein alignment models from the sequence database. Nucleic Acids Res 1997;25:1665–1677.
24. Fischer D. Hybrid fold recognition: combining sequence derived properties with evolutionary information. Pac Symp Biocomput 2000;119–130.
25. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000;302:205–217.
26. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
27. Rost B, Schneider R, Sander C. Protein fold recognition by prediction-based threading. J Mol Biol 1997;270:471–480.
28. Holm L, Sander C. Touring protein fold space with Dali/FSSP. Nucleic Acids Res 1998;26:316–319.
29. Irmler M, Thome M, Hahne M, Schneider P, Hofmann K, Steiner V, Bodmer JL, Schroter M, Burns K, Mattmann C, Rimoldi D, French LE, Tschopp J. Inhibition of death receptor signals by cellular FLIP. Nature 1997;388:190–195.
30. Hu S, Vincenz C, Ni J, Gentz R, Dixit VM. I-FLICE, a novel inhibitor of tumor necrosis factor receptor-1- and CD-95-induced apoptosis. J Biol Chem 1997;272:17255–17257.
31. Koonin EV, Aravind L. The NACHT family—a new group of predicted NTPases implicated in apoptosis and MHC transcription activation. Trends Biochem Sci 2000;25:223–224.
32. Leonard CJ, Aravind L, Koonin EV. Novel families of putative protein kinases in bacteria and archaea: evolution of the "eukaryotic" protein kinase superfamily. Genome Res 1998;8:1038–1047.
33. Gronewold TM, Kaiser D. The act operon controls the level and time of C-signal production for Myxococcus xanthus development. Mol Microbiol 2001;40:744–756.
34. Brown JR, Doolittle WF. Archaea and the prokaryote-to-eukaryote transition. Microbiol Mol Biol Rev 1997;61:456–502.
35. Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, Koonin EV. Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. Genome Res 1999;9:608–628.
36. Doolittle WF. Phylogenetic classification and the universal tree. Science 1999;284:2124–2129.
37. Sogin M. History assignment: when was the mitochondrion founded? Curr Opin Genet Dev 1997;7:792–799.
38. Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. Nature 1998;396:133–140.
39. Doolittle WF. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet 1998;14:307–311.
40. Hasegawa M, Kishino H, Saitou N. On the maximum likelihood method in molecular phylogenetics. J Mol Evol 1991;32:443–445.
41. Baldauf SL, Doolittle WF. Origin and evolution of the slime molds (Mycetozoa). Proc Natl Acad Sci U S A 1997;94:12007–12012.
42. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
43. Kishino H, Miyata T, Hasegawa M. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J Mol Evol 1990;31:151–160.
44. Felsenstein J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol 1996;266:418–427.
45. Swofford DL. Phylogenetic analysis using parsimony and other methods. Sunderland (MA): Sinauer; 1998.
46. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 1997;10:1–6.
47. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714–2723.
48. Kraulis PJ. Molscript. J Appl Crystallogr 1991;24:946–950.