Extensive Expansion of A1 Family Aspartic Proteinases in Fungi Revealed by Evolutionary Analyses of 107 Complete Eukaryotic Proteomes

María V. Revuelta[1], Jan A.L. van Kan[2], John Kay[3], and Arjen ten Have[1,*]

[1]Instituto de Investigaciones Biológicas-CONICET, Universidad Nacional de Mar del Plata, Mar del Plata, Argentina

[2]Laboratory of Phytopathology, Wageningen University, The Netherlands

[3]School of Biosciences, Cardiff University, United Kingdom

*Corresponding author: E-mail: tenhave.arjen@gmail.com, atenhave@mdp.edu.ar.

Accepted: May 21, 2014

Abstract

The A1 family of eukaryotic aspartic proteinases (APs) forms one of the 16 AP families. Although one of the best characterized families, the recent increase in genome sequence data has revealed many fungal AP homologs with novel sequence characteristics. This study was performed to explore the fungal AP sequence space and to obtain an in-depth understanding of fungal AP evolution. Using a comprehensive phylogeny of approximately 700 AP sequences from the complete proteomes of 87 fungi and 20 nonfungal eukaryotes, 11 major clades of APs were defined of which clade I largely corresponds to the A1A subfamily of pepsin-archetype APs. Clade II largely corresponds to the A1B subfamily of nepenthesin-archetype APs. Remarkably, the nine other clades contain only fungal APs, thus indicating that fungal APs have undergone a large sequence diversification. The topology of the tree indicates that fungal APs have been subject to both "birth and death" evolution and "functional redundancy and diversification." This is substantiated by coclustering of certain functional sequence characteristics. A meta-analysis toward the identification of Cluster Determining Positions (CDPs) was performed in order to investigate the structural and biochemical basis for diversification. Seven CDPs contribute to the secondary structure of the enzyme. Three other CDPs are found in the vicinity of the substrate binding cleft. Tree topology, the large sequence variation among fungal APs, and the apparent functional diversification suggest that an amendment to update the current A1 AP classification based on a comprehensive phylogenetic clustering might contribute to refinement of the classification in the MEROPS peptidase database.

Key words: aspartic protease, phylogeny, molecular evolution, functional redundancy and diversification, classification, structure–function prediction.

Introduction

Aspartic proteinases (APs) form one of the eight major classes of protein-degrading enzymes and are divided into five clans or superfamilies (AA, AC, AD, AE, and AF) and 16 families. The AA clan includes the A1 family of eukaryotic APs, which contains many biochemically characterized enzymes. APs of the A1 family are believed to have evolved through gene duplication (Tang 1979). The polypeptide characteristically consists of two internally homologous domains, each of which provides a catalytic Asp residue, positioned within the hallmark motif Asp-Thr/Ser-Gly, to the active site. Each of these Asp residues is followed by a hydrophobic–hydrophobic–Gly sequence. Together, these residues form a structural feature known as a psi-loop, named after its physical resemblance to the Greek letter Ψ (Cooper et al. 1990). The N-terminal domain also contains the strictly conserved Tyr75 residue, located in a β-hairpin loop which overhangs the active site, and a Trp residue commonly, but not absolutely, conserved at position 39 (residue numbering according to that of pig pepsin).

APs fulfill a variety of physiological roles and, currently, the A1 family is further subdivided into subfamilies A1A (pepsin-archetype APs) and A1B (nepenthesin-archetype APs) (Rawlings et al. 2012). Subfamily A1B APs are distinguished from A1A APs by the presence of the nepenthesin-specific insert, a Cys-rich segment inserted between Trp39 and Tyr75 in the polypeptide. Other functionally important inserts

such as the Plant Specific Insert in phytepsins have been identified (Runeberg-Roos et al. 1991) and recently a comparative genomics study of APs in Oomycetes demonstrated that a substantial number of APs have sequence characteristics, such as potential membrane spanning stretches, that suggest different functions (Kay et al. 2011).

Fungi have a relatively high complement of APs that show a large sequence diversity and interspecific clustering (Ten Have et al. 2004, 2010; Monod et al. 2011; Xiao et al. 2011), indicating that the AP gene family in fungi has undergone a process of birth and death rather than concerted evolution (Eirín-López et al. 2012). Fungal-secreted APs used in industrial processes such as aspergillopepsin (Cho et al. 2001), mucorpepsin (Newman et al. 1993), and penicillopepsin (Fraser et al. 1992) have been well-characterized in biochemical terms but little is known about their physiological functions. Fungal APs can also be located in vacuoles or associated with the cell membrane or cell wall through a glycosylphosphatidylinisotol (GPI)-anchor. A recent fungal AP phylogeny (Monod et al. 2011) showed that 21 major clades could be selected. However, the phylogeny did not include complete proteomes from Basidiomycetes or lower fungi, nor sequences from nonfungal eukaryotes. Recently, many complete proteomes from Basidiomycetes, a Chytridiomycete, and other minor taxonomical groups have become public, which enables the reconstruction of a comprehensive phylogeny and which is fundamental in understanding processes such as birth and death evolution (Lemberg and Freeman 2007; Bondino et al. 2012; Castro et al. 2012; Eirín-López et al. 2012). Birth and death evolution promotes genetic diversity and results in functional redundancy and diversification. Combining phylogenetic clustering with biocomputational analysis can give indications for how and in which evolutionary lineages functional diversification has occurred. The cellular location to which a polypeptide is directed is an important functional aspect and can be inferred from sequence motifs. Biocomputational tools have been developed in order to identify "Cluster Determining Positions" (CDPs), which correspond to residues that are important for the evolution of a protein superfamily. Analysis of the residues at CDPs, using both 3D structures and high quality multiple sequence alignments (MSAs), can provide insight into functional aspects of the family members. Such studies can identify APs with novel interesting features and validate classifications derived from the hierarchical clustering of the phylogeny.

Based on the above considerations, this study was performed in order to explore the fungal AP sequence space, to obtain a more profound understanding of fungal AP evolution, to obtain indications for functional diversification of fungal APs, and to explore whether MEROPS might benefit from an amendment of the current classification of the A1 family of eukaryotic APs.

Materials and Methods

Data Resources

The 107 complete proteomes were downloaded from various sources, as specified in supplementary table S1*A*, Supplementary Material online. The PDB database was downloaded from National Center for Biotechnology Information (NCBI)'s FTP site (ftp://ftp.ncbi.nih.gov/blast/db/FASTA/pdbaa.gz, last accessed June 13, 2014).

Identification of Functional AP Homologs

Separate HMMER profiles were constructed for the MEROPS A1A and A1B alignments by means of "hmmbuild" using default settings (HMMER Version 3.0; Eddy 2011). The complete proteomes of 107 organisms (see fig. 1*B* and *C* and supplementary table S1*A*, Supplementary Material online) as well as the PDB database were then screened by "hmmsearch" using the two profiles and default settings. All sequences identified by the matrices with a full sequence *E* value smaller than the HMMER exclusion threshold were considered as AP homologs. First and for identification purposes only, the N- and C-terminal regions of the sequences were trimmed according to the MSA published by Ten Have et al. (2010). Basically, this consists in the removal of nonhomologous extensions beyond the mature polypeptide, which would generate noise in subsequent sequence filtering. Then, the sequence set was reduced in size by means of CD-HIT using 85% identity as upper threshold (Huang et al. 2010). The remaining set was scrutinized for the presence of typical AP hallmarks defined as D^{32}[TS]G, Y^{75}, XXG^{122}, D^{215}[TS]G, and XXG^{302} (where X is any of the hydrophobic residues AFILMV and numbering is according to mature pig pepsin). Sequences lacking any of the hallmarks were considered nonfunctional homologs and excluded from further analysis. A second screening was performed directed at the lack of secondary structure elements. Sequences and secondary structure information obtained from the AP structures were mapped onto a preliminary MSA and sequences that appeared to lack important β sheets were removed. Then, sequences with long (>15 amino acids) inserts were removed that appeared to have no homologous counterparts in any of the sequences collected or in a homologous sequence identified by BLAST in the nonredundant database of NCBI.

MSA and Phylogenetic Analysis

Multiple protein sequence alignments were performed using the PROMALS3D program (Pei et al. 2008). Trimming for phylogeny was performed with Block Mapping and Gathering with Entropy (BMGE) (Criscuolo and Gribaldo 2010) using the command options "-t AA -m BLOSUM30 -b 1 -g 0.9." Maximum likelihood phylogenies were built using PhyML-a-bayes (Guindon et al. 2010). Specifically, PhyML analyses were conducted with the LG model, estimated proportion

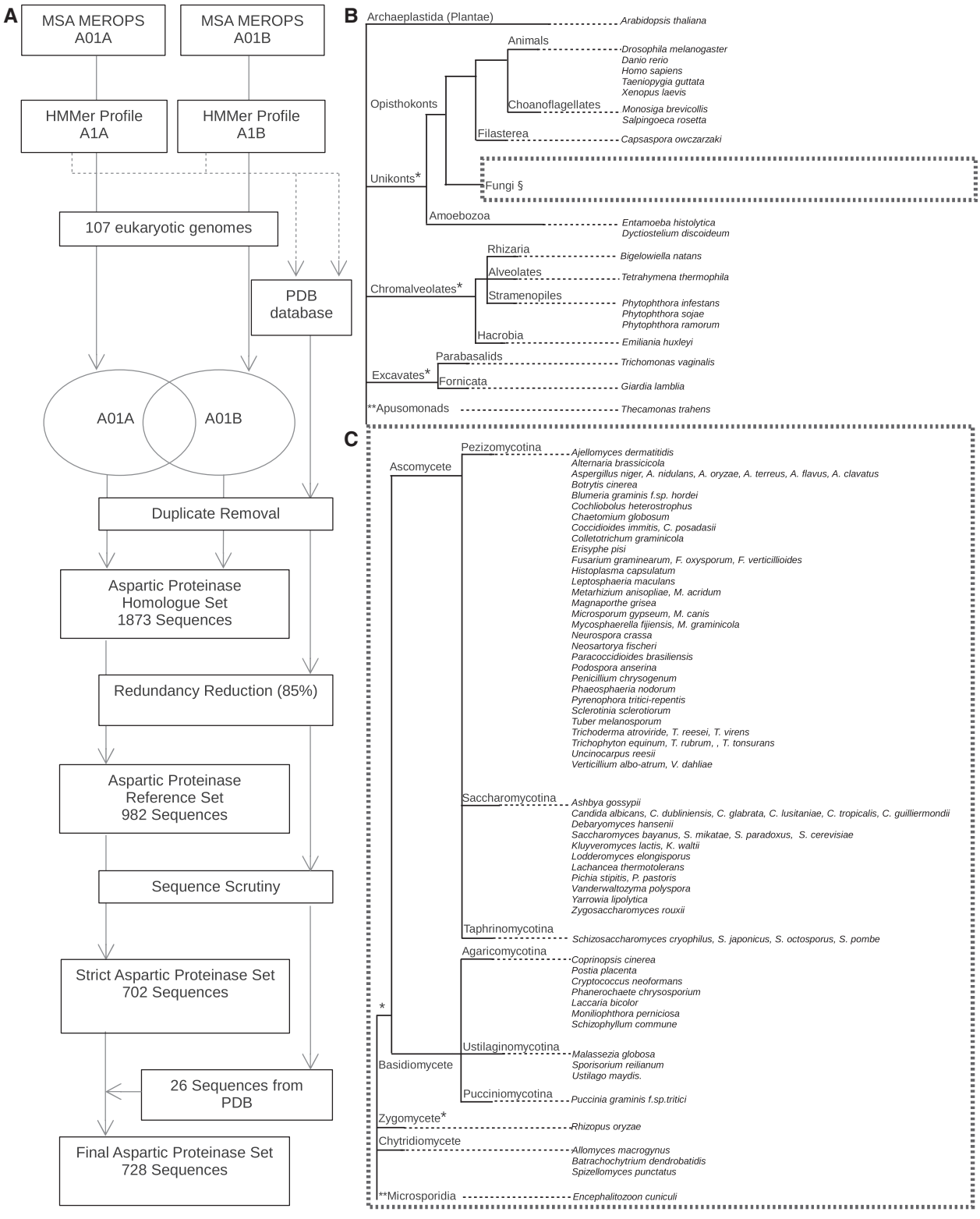

Fig. 1.—Sequence mining flowchart and taxonomic sequence sampling of APs. (*A*) HMMER profiles built from the MEROPS Peptidase Database A1A (pepsin-archetype) and A1B (nepenthesin-archetype) holotype alignments were used to screen both 107 eukaryotic complete proteomes and the Protein Data Bank (PDB) database. Consecutive steps of information redundancy reduction and sequence hallmark scrutiny were performed in order to achieve the final set of 728 AP sequences including 26 sequences for which a 3D structure is available. (*B*) Taxonomic organization of the 107 genomes examined. *These clades may not be monophyletic § see panel (*C*) Detail of fungal sequence sampling. **Placement of this clade is without consensus.

of invariable sites, four rate categories, estimated gamma distribution parameter, and optimized starting BIONJ tree, both with aLRT and 100 bootstraps branch support measures. The resulting phylogenetic trees were viewed and edited with iTol version 2.0 (Letunic and Bork 2011), Dendroscope (Huson and Scornavacca 2012), and Inkscape (GNU license, www.inkscape.org, last accessed June 13, 2014).

Additional Biocomputational Analyses

The SignalP server (Petersen et al. 2011) (http://www.cbs.dtu.dk/services/SignalP/, last accessed June 13, 2014) was used to predict the presence of signal peptides, using complete sequences. Prediction of nonclassical secretion signals was performed using the SecretomeP server (Bendtsen et al. 2004) (http://www.cbs.dtu.dk/services/SecretomeP/, last accessed June 13, 2014), trained with mammalian sequences, due to the lack of fungal-specific prediction software. WolfPSort (Horton et al. 2007) was used to determine putative subcellular localization. The prediction of GPI-motifs was performed using the web-tool bigPI fungal predictor (Eisenhaber et al. 2003) (http://mendel.imp.ac.at/gpi/fungi_server.html, last accessed June 13, 2014). Prediction of transmembrane helices was made with TMHMM server (Krogh et al. 2001) (http://www.cbs.dtu.dk/services/TMHMM/, last accessed June 13, 2014).

For the CDP analyses, three different software packages were used. SDPfox (Mazin et al. 2010) (http://bioinf.fbb.msu.ru/SDPfoxWeb/main.jsp, last accessed June 13, 2014) was used with default parameters, allowing a maximum of 30% of gaps in a group per column. GroupSim (Capra and Singh 2008) was executed locally, with the following parameters: -w 0 -m BLOSUM30.txt. GroupSim CDPs were selected when having a score higher than 1.5 (total score achieved = 3.7). Coupled Mutation Finder (CMF) (Haubrock et al. 2012) was performed online (cmf.bioinf.med.uni-goettingen.de/, last accessed June 13, 2014) with default parameters, using the U-metric and the UD(α)-metric.

Sequence LOGOS were created with WebLogo (Crooks et al. 2004) (http://weblogo.berkeley.edu/, last accessed June 13, 2014), and AP structures were visualized and edited with Jmol (www.jmol.org, last accessed June 13, 2014).

Results

Identification of AP Sequences in 107 Genomes

The strategy to identify and collect AP-encoding sequences was based on a number of considerations. A comprehensive analysis of fungal APs should be embedded in a more expansive taxonomical context. Then, similar sequences that lack strictly conserved hallmark residues should be removed as they might not correspond to functional APs and therefore are under a different functional constraint. Inadvertently, a number of real positives (i.e., sequences lacking hallmark residues that nevertheless encode for functional APs) might be removed in order to prevent inclusion of false positives (i.e., similar sequences that do not encode functional APs). Two additional technical aspects were considered. As structure is more conserved than sequence, inclusion of sequences of APs for which X-ray structures have already been resolved would enhance the quality of the MSA; and, in order to facilitate computation of the alignments and phylogenies, redundancy of information should be avoided. We devised the data mining protocol depicted in figure 1*A* to screen 87 complete fungal proteomes (fig. 1*C*) along with 20 complete proteomes from nonfungal eukaryotes (fig. 1*B*), both sets being selected as a representative sample with a broad taxonomic distribution (see also supplementary table S1*A*, Supplementary Material online, for detailed information).

Two HMMER profile matrices (Eddy 2011) were constructed using the A1A and A1B subfamily alignments retrieved from the MEROPS database (Rawlings et al. 2012). When the HMMER profiles obtained were used together to screen the 107 complete proteomes, the sequences of 1,873 AP homologs were identified. In order to reduce information redundancy, sequences with >85% identity to any other sequence were removed using CD-HIT (Huang et al. 2010). The resulting set was then scrutinized for the presence of the AP hallmark motifs (see Introduction) that were defined as D^{32}[TS]G, Y^{75}, XXG^{122}, D^{215}[TS]G, and XXG^{302} (where X is any hydrophobic residue and numbering is according to mature pig pepsin). Sequences of 26 APs with determined X-ray structures were retrieved from the PDB database (supplementary table S1*B*, Supplementary Material online) (Berman et al. 2000) and added to the reduced, scrutinized sequence set, yielding a final set of 728 sequences. This subset of sequences, from which nonhomologous N- and C- terminal extensions were removed (according to Ten Have et al. [2010]), is available in supplementary datafile S1 (fasta), Supplementary Material online.

Alignment Construction and Phylogenetic Clustering of APs

An MSA of the 728 AP sequences was constructed by means of PROMALS3D (Pei et al. 2008) followed by manual correction. An excerpt of the MSA, with strategically chosen sequences for which 3D structures have all been resolved, is depicted in figure 2. β-Sheets, α-helices, and the aforementioned AP sequence hallmarks are highlighted and appear to be correctly aligned.

As the 728 sequences are very divergent and represent different evolutionary histories, the MSA contains blocks of poorly aligned subsequences including nonhomologous or cluster-specific subsequences such as the nepenthesin-specific insert (see Introduction). These were removed by BMGE (Criscuolo and Gribaldo 2010), which permits selection of parts of the alignment that are suitable for proper

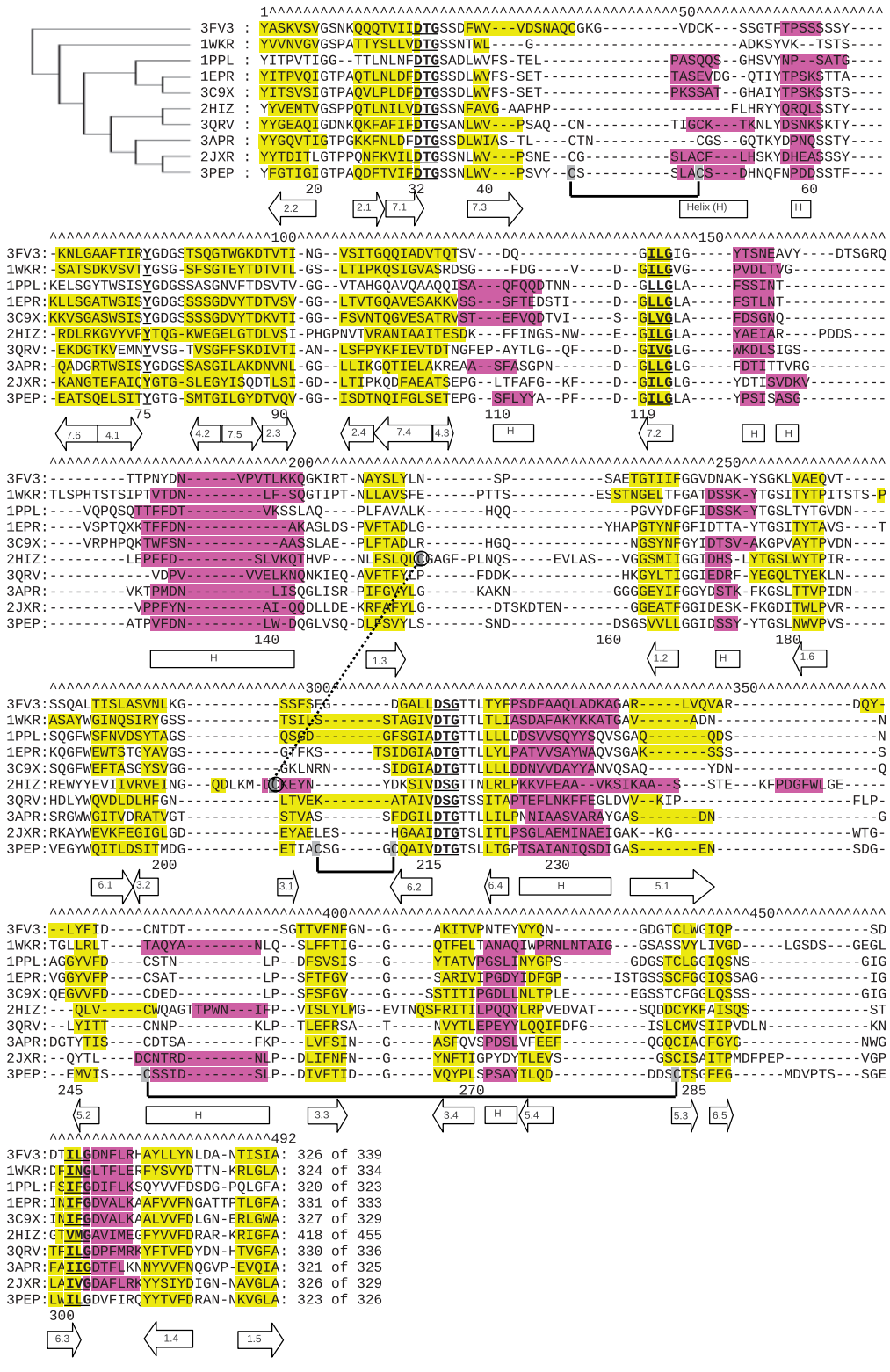

Fig. 2.—Excerpt of MSA of APs. The aligned amino acid sequences are a representative subset of ten APs for which structures are known, indicated by their PDB access codes, that have been extracted from the main alignment. The tree at the top left-hand corner demonstrates the phylogenetic relationship between the sequences as derived from the main phylogeny (fig. 3). The standard numbering for the mature region of pig pepsin (3PEP) is included underneath the sequence blocks whereas numbers above correspond to the alignment columns. The characteristic D[TS]G and hydrophobic–hydrophobic–Gly motifs forming the psi loops in each domain of APs as well as the strictly conserved Tyr75 are underlined in bold. The pairing of the Cys residues forming

(continued)

phylogenetic inference. The trimmed alignment is presented in supplementary datafile S2 (fasta), Supplementary Material online. The trimmed MSA was used to reconstruct a phylogenetic tree using PHYML-a-Bayes (Guindon et al. 2010). As shown in figure 3, 11 major monophyletic clades were assigned. Nine sequences did not cluster in any of these monophyletic clades.

Two novel deductions can be made from the phylogenetic tree and the clustering pattern. First, with the exception of two orphans (sequences from *Emiliania huxleyi* and *Monosiga brevicollis* that appear at a long distance in clade XI), all APs from organisms outside the Kingdom Fungi fall into two clades. Clade I contains only sequences that correspond to the A1A subfamily whereas clade II contains all sequences that correspond to A1B, supplemented with a number of sequences from the A1A subfamily (as determined by MEROPS BLAST). Hence, this part of the phylogeny largely correlates with the MEROPS classification into pepsin-archetype (A1A) and nepenthesin-archetype (A1B). Fungal AP sequences are also present within clades I and II. The last common ancestor (LCA) of all APs is therefore most likely found at the node that connects clades I and II (see fig. 3*B*). The other nine clades consist exclusively of fungal sequences, with the exception of the two aforementioned orphans. Many of these clades contain sequences that in the MEROPS database appear as an A1A AP. Our phylogenetic clustering would thus appear to provide greater detail and a higher resolution than MEROPS classification. Second and remarkably, the pepsin- and nepenthesin-archetype APs appear to encompass only a small part of AP sequence space as clades I and II form a monophyletic superclade in a tree with nine other clades that are more distant. The distances between the 11 clades are most readily visualized through the radial phylogram of figure 3*B*.

Separate phylogenies for all 11 major clades were reconstructed based on the corresponding subMSAs. As these subMSAs contain less divergent sequences, BMGE selected more columns for all subsequent phylogenies thus increasing the resolution of the phylogeny. The main results are discussed below. All trees are available in supplementary datafile S3 (txt), Supplementary Material online.

Clade I: The "Pepsin" Archetype APs

Clade I of figure 3, depicted at higher resolution in figure 4*A*, contains only sequences classified by MEROPS as A1A and is therefore referred to as the pepsin-archetype clade. It contains APs involved in a diverse range of processes like digestion and blood pressure control. Gastric enzymes such as pepsin, gastricsin, and chymosin (for PDB details, see supplementary table S1*B*, Supplementary Material online) cluster in a small subclade. Another subclade with only metazoan sequences contains lysosomal cathepsin D, napsin and the AP involved in the regulation of blood pressure, renin. A large subclade is formed by vacuolar enzymes from fungi such as proteinase A from *Saccharomyces cerevisiae* (Parr et al. 2007). All fungi, except for the Zygomycete *Rhizopus oryzae* and Taphrinomycetes, have a single ortholog in this cluster. *Rhizopus oryzae* has two homologs (probably as the result of a recent whole genome duplication [Ma et al. 2009]), whereas Taphrinomycetes lack an ortholog. These vacuolar fungal APs cluster with representatives of the chromalveolata such as PiAP1 from *Phytophthora infestans*. The topology of the vacuolar subclade is largely consistent with the topology of the species tree (fig. 1*C*). Nonvacuolar, Zygomycete APs, among which is Rhizopuspepsin, form a small polyphyletic group. The two remaining subclades contain, respectively, sequences from a number of taxonomic domains including oomycetes (PsAP4) and plants (Phytepsin).

Clade II: The "Nepenthesin"-Archetype APs and Close Homologs

Clade II (see fig. 4*B*) contains the nepenthesin-archetype or A1B APs that could be considered as true nepenthesin homologs, as well as a number of representatives from the MEROPS A1A subfamily. Nepenthesins are named after APs that were first isolated from the slender-pitcher plant *Nepenthes gracilis* (Takahashi et al. 2005). True nepenthesin homologs have an insert between Trp39 and Tyr75 (see Introduction). This insert contains a number of cysteine residues, contributing additional disulphide bridges which have been suggested to account for the enhanced stability of nepenthesin over a wide pH range (Athauda et al. 2004; Takahashi et al. 2005). In clade II, all of the nepenthesins assigned to the MEROPS A1B subfamily cluster within a subclade (labeled as the nepenthesin subclade in fig. 4*B*). Within this subclade, the nepenthesins from *Arabidopsis thaliana* cluster into a monophyletic sub-subclade whereas other nepenthesins that are not of plant origin such as AP10, AP11 and AP12 from *Phytophthora* species fall into a separate sub-subclade. The latter also contains a number of sequences hitherto classified as pepsin-archetype (A1A) and which all lack the Cys-rich insert. Among the latter are other *Phytophthora* APs (AP2 and AP8) and also human BACE1 and BACE2 (fig. 4*B*). Another subclade with sequences that have been classified

Fig. 2.—Continued

the three disulphide bonds in pig pepsin is indicated by thick lines. The Cys residues forming an additional disulphide bridge in human BACE1 (2HIZ) are indicated by a circle and a thick dotted line. β-Sheets and α-helices are highlighted by yellow and magenta shading, respectively. The exact organization of secondary structures found in pig pepsin is depicted below the alignment.

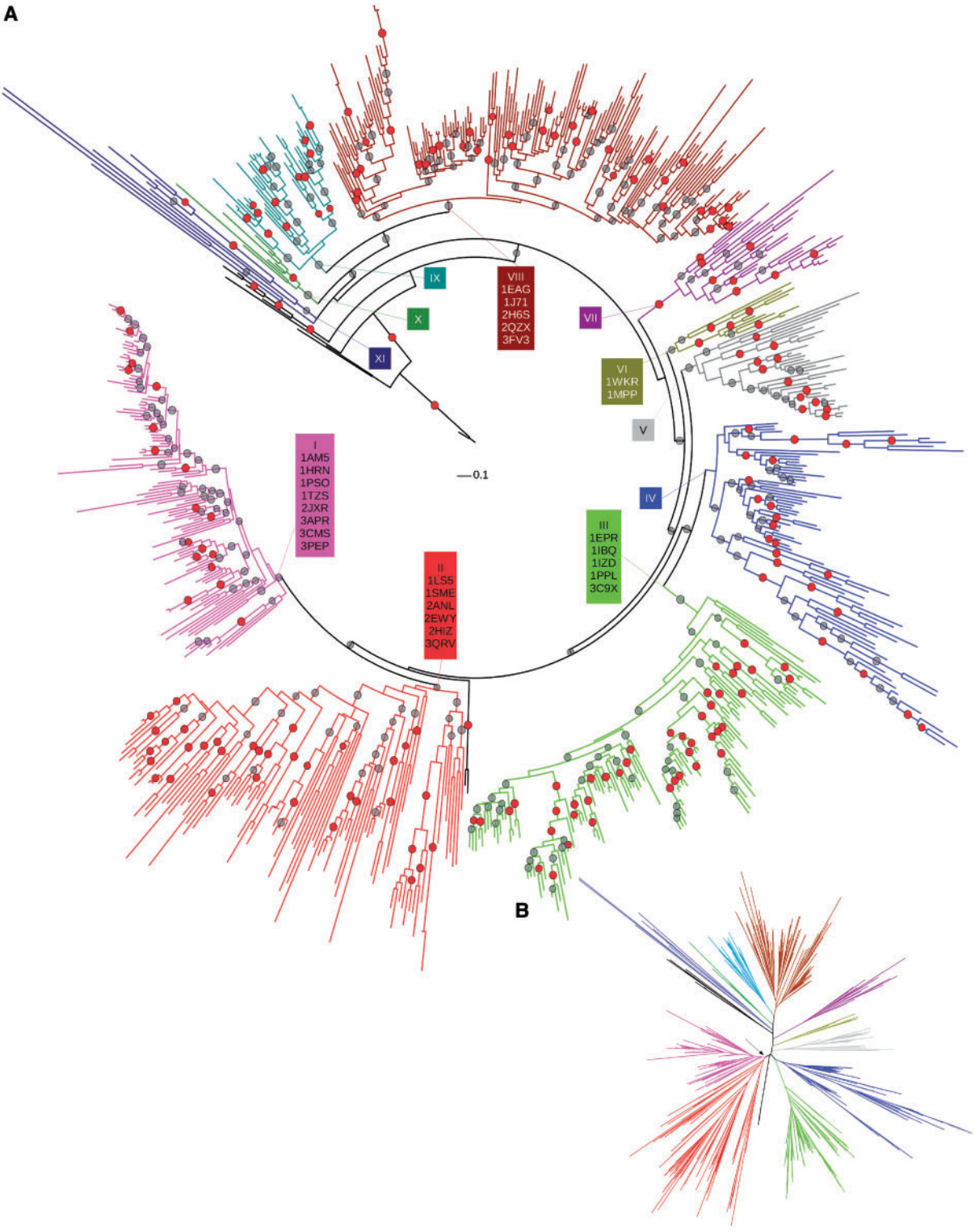

Fig. 3.—Phylogenetic tree of eukaryotic APs. The 728 AP sequences were aligned, trimmed, and subjected to phylogenetic analysis by maximum likelihood using PHYML-a-Bayes. (*A*) Circular phylogram in which each of the 11 monophyletic clades is depicted in a different color, with color-matched boxes indicating the clade number and the PDB structures contained therein. Red dots placed on edges indicate both aLRT and bootstrap support of ≥80%, grey dots placed on edges indicate ≥80% aLRT support only. Nine orphan sequences are indicated in black. The scale bar indicates 0.1 amino acid substitution per site. (*B*) Radial phylogram. Colors as in (A), magenta and red clades correspond to pepsin and nepenthesin-archetype APs whereas the other clades contain only fungal AP sequences. The arrow points to the most probable LCA of A1 APs.

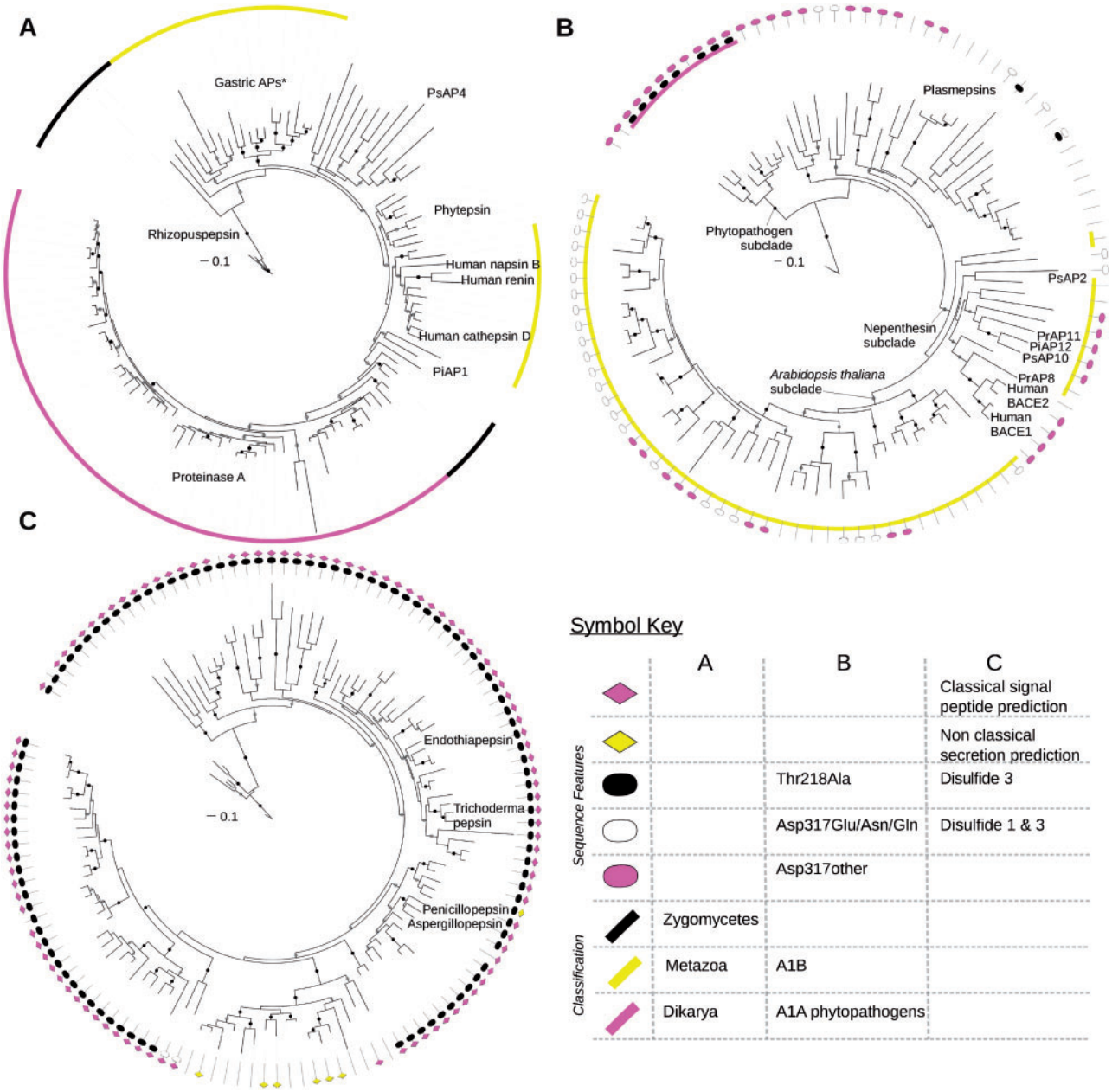

Fig. 4.—Pepsin, nepenthesin, and clade III AP phylogenetic trees. Subsets of the aligned sequences corresponding to clades I, II, and III from the main phylogenetic tree (fig. 3) were independently trimmed and used for phylogenetic analysis by maximum likelihood using PHYML-a-Bayes. Black dots placed on edges indicate both aLRT and bootstrap support of ≥80%, grey dots placed on edges indicate ≥80% aLRT support only. The scale bars indicates 0.1 amino acid substitution per site. (*A*) Pepsin archetype APs, (*B*) nepenthesin APs and their homologs, (*C*) clade III fungal APs. Classifications and particular amino acid sequence features are annotated on the surrounding rings according to the embedded symbol key. *Gastric APs include pig, human and atlantic cod pepsins, human gastricsin, calf chymosin, and human cathepsin E.

previously as belonging to the A1A subfamily, such as plasmepsins I, II and IV from the vacuoles of *Plasmodium* spp, also contains sequences from the genus *Phytophthora*, from *Tetrahymena*, from three higher fungi and two minor clades of the lower Chytridiomycete fungi. A further monophyletic subclade consists only of nine APs from higher fungi, all of which are phytopathogens (labeled as phytopathogen subclade, fig. 4*B*). Additionally, all of the sequences in this subclade are typified by the absence of a propeptide and a substitution of the almost invariant Asp317 residue by Ala, Thr, Pro or Met; in addition eight out of the nine sequences carry a substitution of the commonplace Thr/Ser218 residue by Ala (fig. 4*B* and supplementary datafile S2, Supplementary Material online).

Secreted Fungal APs

Clade III (see fig. 4*C*) contains only sequences from the fungal phyla Ascomycetes and Basidiomycetes, and includes the well-described aspergillopepsin, penicillopepsin, trichodermapepsin and edindothiapepsin, for which 3D structures have long been determined. Some of these enzymes have been demonstrated experimentally to be secreted. Indeed, most entries in figure 4*C* are predicted to have signal peptides. However, the sequences of a small subclade of APs appear to lack a signal peptide, according to both WolfPSort (Horton et al. 2007) and SignalP 4.0 (Petersen et al. 2011) servers. This lack of a signal peptide coclusters with the complete absence of disulfide bridges in these AP sequences. A number of the APs in this subclade might therefore be either cytosolic or subject to secretion by a nonclassical pathway. BcAP1 from *Botrytis cinerea*, a member of this particular subclade, was identified in the secretome in one study, but the authors comment it might be due to contamination (Li et al. 2012). Other secretome studies on *B. cinerea* did identify a number of APs but not BcAP1 (Shah et al. 2009; Espino et al. 2010; Fernández-Acero et al. 2010).

Four other clades, VI, VII, X and XI, contain only fungal sequences that correspond with secreted APs but with no other obvious distinctive characteristics. Clade VI includes the sequences with determined structures of polyporopepsin and mucorpepsin as well as the recently described BcAP8, which is the most abundant secreted proteinase of *B. cinerea* (Ten Have et al. 2010).

Cell Wall Associated APs and Their Close Homologs

Clades IV, V, VIII, and IX consist mostly of sequences with a C-terminal extension. Some C-terminal subsequences of clade IV are predicted to act as transmembrane regions, whereas others contain a GPI-anchor motif. The presence of predicted transmembrane regions and GPI-anchor motifs largely follows the hierarchical clustering of clade IV. Clade IV (supplementary fig. S1*A*, Supplementary Material online) consists of Basidiomycete sequences only. All Basidiomycete fungi included in this study, except *Moniliophthora perniciosa,* have at least one homolog in this group. The sequences in clade V are derived from both Ascomycetes and Basidiomycetes, and mostly encompass APs with GPI-anchor motifs (supplementary fig. S1*B*, Supplementary Material online). To the best of our knowledge, clades IV and V do not contain members that have been biochemically characterized.

Clade VIII (fig. 5) contains the yapsins, a group of GPI-anchored APs (and close homologs such as barrierpepsin) that have been extensively characterized and reported to be involved in various processes such as the maturation of secreted hydrolases, pathogenesis (Monod et al. 2002; Albrecht et al. 2006), pheromone proteolysis, sexual reproduction (Alby et al. 2009), and cell wall integrity (Gagnon-Arsenault et al. 2006). The yapsin homologs of clade VIII

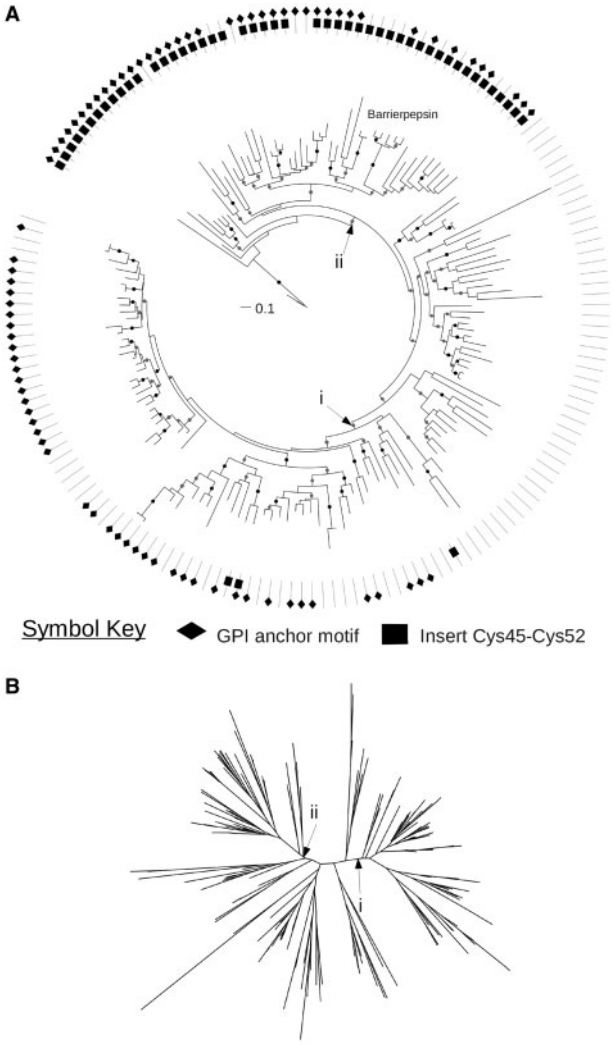

Fig. 5.—Phylogenetic tree of yapsin and yapsin-like APs. The aligned sequences corresponding to clade VIII from the main phylogenetic tree (fig. 3) were separately trimmed and used for phylogenetic analysis by maximum likelihood using PHYML-a-Bayes. (*A*) Circular phylogram. Black dots placed on edges indicate both aLRT and bootstrap support of ≥80%, grey dots placed on edges indicate ≥80% aLRT support only. The scale bar indicates 0.1 amino acid substitution per site. Presence of GPI-anchor motif and the insert between Cys45 and Cys52 are annotated on the surrounding rings according to the embedded symbol key. (*B*) Radial phylogram: Monophyletic subclade i corresponds to Pezizomycete APs, monophyletic subclade ii corresponds to Saccharomycete APs with GPI-anchor and a cluster specific subsequence between Cys-45 and Cys-52. The other subclades correspond with Saccharomycete APs that lack these cluster-specific subsequences.

were detected only in Saccharomycetes and Pezizomycetes and not in any of the other analyzed taxonomic groups. Most of the Pezizomycete sequences have GPI-anchor motifs in the C-terminal extensions (monophyletic subclade i, fig. 5). Approximately half of the Saccharomycete APs also

A

Pos	C-U	C-UD	Group	SDP	SecS
37		x	x	x	No
41	x		x	x	VII-3
84	x		x	x	VII-5
129	x	x	x	x	No
151		x	x	x	I-3
153	x		x	x	I-3
165	x		x	x	I-6
232	x		x	x	H-8
284	x		x	x	No
309	x		x	x	I-4

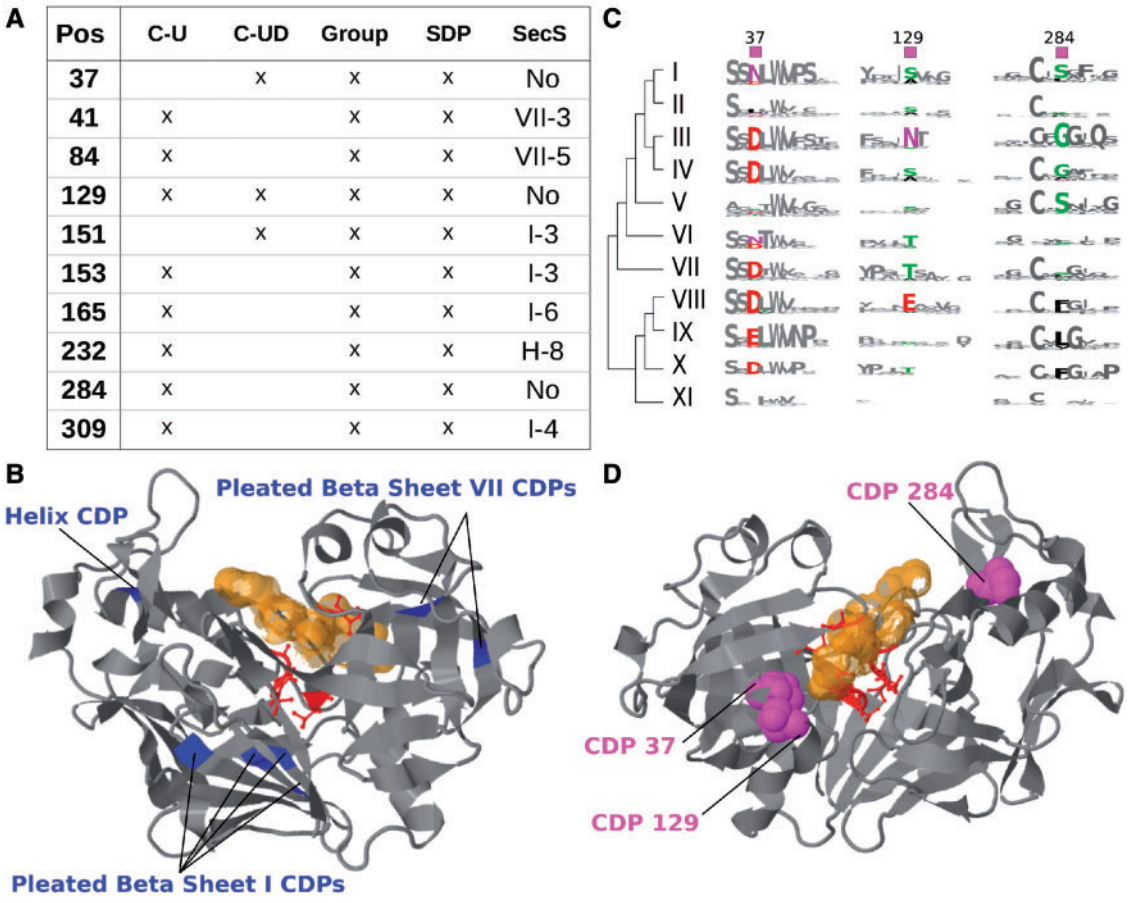

Fig. 6.—CDPs of eukaryotic APs. (*A*) CDP analysis was performed using four separate algorithms (C-U/C-UD, Group and SDP refer to the CMF U/UD metrics, GroupSim, and SDPfox; supplementary table S2, Supplementary Material online), whereas SecS locates each CDP in the secondary structure of human pepsin (PDB structure 1PSO; H, I and VII refer to helix, pleated beta-sheets I and VII, respectively). Residues (numbered according to the sequence of human pepsin) identified by at least three of these methods were considered to be CDPs. (*B*) Cartoon of structure of the complex between human pepsin and the inhibitor pepstatin (in yellow). The catalytic D[TS]G motifs and Y75 residues are all depicted (in red) together with CDPs 41, 84, 151, 153, 165, 232, and 309 contributing to helices and B-pleated sheets (in blue). (*C*) Sequence logos of subsequences flanking three selected CDPs (37, 129, and 284) are depicted for all major clades (fig. 3) with colors according to standard physicochemical characteristics. (*D*) Cartoon showing the pepsin–pepstatin structure viewed from a different angle from that in (*B*) with CDPs 37, 129, and 284 (in pink, 100% van der Waals) adjacent to the active site.

have a GPI-anchor motif and cluster in a monophyletic subclade ii (fig. 5*B*), which includes the GPI lacking barrierpepsin. Monophyly is substantiated by the presence of a cluster specific subsequence between Cys45 and Cys52 that was identified earlier (Abad-Zapatero et al. 1996) (fig. 5*A*). Intriguingly however, a GPI-anchor motif is lacking in the rest of the sequences from the Saccharomycetes. It should be noted that most Saccharomycete species sampled for this study have a homolog in both clades, which indicates that a duplication of the yapsin gene has occurred in the LCA of the Saccharomycetes.

Clade IX is monophyletic with clade VIII, and contains exclusively Pezizomycete sequences, most of which also have a C-terminal extension often containing a GPI-anchor motif (data not shown).

Cluster Determining Positions

Software tools (De Juan et al. 2013) have been developed to identify columns in MSAs—or positions in protein structures—that show a high level of conservation within, but not necessarily between, clusters. Such a conservation pattern suggests that these CDPs are structurally or functionally important for the evolution of protein families.

SDPfox (Mazin et al. 2010), Groupsim (Capra and Singh 2008) and CMF (Haubrock et al. 2012) were used to predict CDPs, the last of which used settings of 1) pure entropy and 2) similarity-corrected entropy. When three out of these four methods detected the same position, it was considered a significant CDP and justified for inclusion in a first explorative evaluation. The results of the four methods are collected in supplementary table S2, Supplementary Material online, and

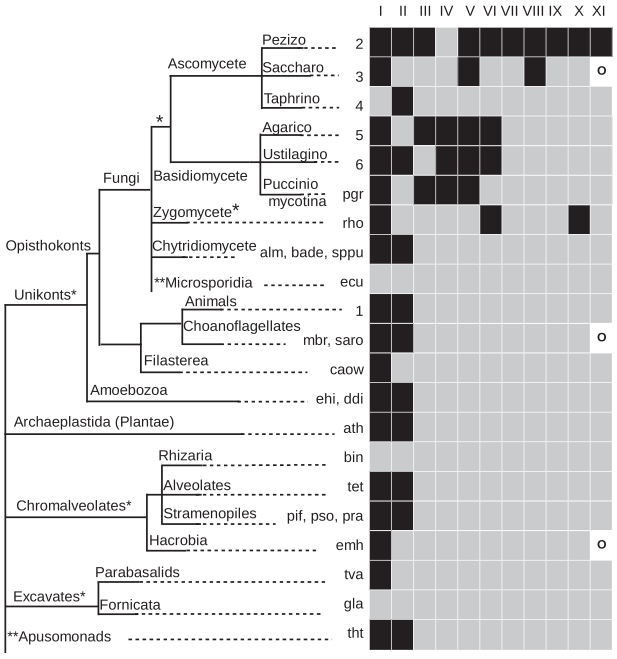

Fig. 7.—Taxonomic distribution of 11 major classes of eukaryotic APs. The occurrence of APs from each of the 11 monophyletic clades across the indicated taxonomic phyla is indicated by black shading, orphans are indicated by an o. *These clades may not be monophyletic; **placement of this clade is without consensus; 1) dme, dre, hsa, tgu, xla; 2) act, afv, ajd, alb, ang, ani, aor, ast, bc, bgr, che, chg, cim, cog, cpw, erp, fgr, fuo, fuv, hca, lem, mean, meac, mgr, mgy, mic, myf, myg, ncr, nfi, pabr, pan, pcs, pno, pyt, ssl, tml, tra, tre, treq, tru, trv, tto, ure, vaa, ved; 3) ago, cal, cdu, cgr, clu, ctp, dha, dkwa, dsba, dsmi, dsrd, kla, lel, lth, pgu, pic, ppa, sce, vpo, yli, zro; 4) schc, schj, scho, spo; 5) cci, cnb, cne, dpch, lbc, mpr, scm; 6) mgl, sre, uma. Complete names of abbreviated species are in figure 1*B* and supplementary table S1*A*, Supplementary Material online.

the ten CDPs thus identified are shown in figure 6*A*. Seven of the CDPs take part in secondary structure of the core of the protein molecule. CDPs 41 and 84 are part of pleated β-sheet VII; CDPs 151, 153, 165, and 309 form part of pleated β-sheet I; and CDP232 is part of helix 8 (fig. 6*B*). CDPs 37, 129 and 284 do not take part in secondary structure, suggesting that they may be under functional rather than structural constraint. These three CDPs all lie in close juxtaposition to the active site, are partially solvent exposed, and can be envisaged to interact with a substrate or inhibitor (fig. 6*D*). Sequence logos for these CDPs and immediate surrounding subsequences (Schneider and Stephens 1990; Crooks et al. 2004) are depicted in figure 6*C* with residues E37 in clade IX and E129 in clade VIII identified as the most conspicuous residues.

Discussion

A1 APs are ubiquitous in eukaryotes (See fig. 7) and have many different cellular and physiological functions. Although a number of fungal APs have been characterized biochemically, genome sequences uncover many APs with novel sequence features. The objective of this study was to explore the diversity of fungal APs. Phylogeny shows that fungal APs have been subject to birth and death evolution, as well as functional redundancy and diversification, which led to the emergence of different subfamilies. The functions of certain subfamilies, such as the yapsins, have been well established, whereas functions for other subfamilies with distinct sequence features remain to be characterized. The results suggest that an update or amendment of the current MEROPS classification would improve future functional annotation of A1 AP encoding sequences.

The AP phylogeny (figs. 3–5) reflects how AP evolution has occurred, according to the maximum likelihood criterion and a carefully constructed MSA. Sequence mining (fig. 1*A*) was directed at the inclusion of all sequences of functional APs and the exclusion of all sequences of nonfunctional AP-homologs. The taxonomic sampling was focused on fungi (fig. 1*C*) but we deliberately included taxa representative of the major eukaryotic lineages, in order to provide a data set that covers the complete AP sequence space (fig. 1*B*). The set was complemented with 26 sequences for which X-ray structures have been resolved. Promals3D, which uses structural information, was used for MSA construction (MSAs made by other programs were inferior—data not shown). The complexity of the evolution of protein superfamilies (where superfamilies include paralogs as well as orthologs, and show functional diversification), demands stringent data processing. Trimming of the nonhomologous or poorly aligned blocks by BMGE (Criscuolo and Gribaldo 2010) was guided by the structural information. The excerpt of the complete MSA (fig. 2) shows that the hallmark residues and distinct sheets and helices, as indicated by the 3D structures, are generally correctly aligned.

Statistical support for the phylogenetic tree was derived from approximate Rate Likelihood Test (Anisimova and Gascuel 2006) and the more conventional and conservative bootstrap analysis. Significant bootstrap support (set at ≥80%) was substantiated in all cases by significant aLRT support. Some of the major clades did not have significant bootstrap support. However, as most clades have significant aLRT support and bootstrap has been proclaimed to be excessively conservative (Anisimova et al. 2011), the tree is well supported. Other results that support the tree topology are the evident clustering of nepenthesins (fig. 4*B*) (despite the nepenthesin-specific insert having been excluded from the phylogenetic reconstruction); the observation that the phylogeny of vacuolar APs (fig. 4*A*) largely coincides with the species taxonomy (fig. 1*C*); and the finding that certain (sub)clades exclusively contain members of a defined taxonomic group. Additional analyses to evaluate incorrect placement of sequences performed by HMMER profiling with 11 clade-specific profiles, corroborated the topology (data not shown).

Altogether both the MSA and the phylogeny can be considered as comprehensive and reliable, even though APs in certain cases show very low mutual similarity.

Although sequence-based maximum likelihood clustering uses all columns for tree reconstruction, certain columns affect ramification of the tree more than others. These so-called CDPs can contribute to our understanding of evolution and functional diversification. Of the many software tools available for the identification of CDPs, we selected three packages and four methods. The global meta-analysis (fig. 6) identified only a few CDPs, seven of which appeared to be important in maintaining the structure of the core of APs and are therefore likely under a structural constraint, rather than related to functional diversification. Three other CDPs are located in the vicinity of the active site, of which CDP37 and CDP129 show residues with clearly different physicochemical characteristics in one cluster, as discussed below.

AP Gene Duplications in Fungi Resulted in Both "Birth and Death Evolution" and "Functional Redundancy and Diversification"

The complexity of the overall phylogeny (fig. 3) and the derived clade-specific phylogenies (figs. 4 and 5) demonstrates that the A1 AP protein family in fungi has undergone an extensive evolution, as compared with APs in other eukaryotic lineages. Many fungal species have APs in various clades (fig. 7); gene duplication at different evolutionary timepoints has resulted in paralogs becoming dispersed in different clades of the phylogenetic tree. On the other hand, many clades lack homologs from certain taxonomic groups. A notable example occurs within the extensive vacuolar subclade (fig. 4*A*) of clade I. The patriarch of the vacuolar AP subfamily, proteinase A or saccharopepsin from *S. cerevisiae*, is essential for processing of precursors of carboxypeptidase Y and proteinase B within the vacuole (Parr et al. 2007). With the exception of Taphrinomycetes, all of the fungal species examined (as well as the taxonomically distant oomycetes of the *Phytophthora* genus) have a highly conserved AP within this vacuolar subclade (fig. 7). In the model Taphrinomycete, *Schizosaccharomyces pombe*, this apparent loss-of-function of a vacuolar AP is compensated by two serine proteinases, ISP6 and PSP3 (homologs of proteinase B in *S. cerevisiae*) that fulfill the proteolytic function in the vacuole (Mukaiyama et al. 2011). Thus, some genes survive in a lineage and others are lost and fungal AP evolution appears to be subject to a birth and death process. Another example is found in clade VI which contains sequences from Zygomycetes, Basidiomycetes, and Ascomycetes (fig. 7). Clade VI homologs must have emerged early in fungal evolution and the relatively short branches that separate these sequences from their common ancestor indicate that the proteins are under constraint and important for these fungi. Nevertheless, most fungal species lack a representative in this clade, demonstrating that the clade VI homolog is not vital to all fungi. There appears to be no obvious common feature of the fungi that have a homolog in clade VI. As an example, the Ascomycete phytopathogen *B. cinerea* secretes a highly expressed clade VI ortholog (known as BcAP8) whereas the taxonomically closely related phytopathogen, *Sclerotinia sclerotiorum,* lacks this ortholog (Ten Have et al. 2010), despite a similar lifestyle and infection strategy (Amselem et al. 2011).

The LCA of all APs is most likely found at the node that connects clades I and II (see fig. 3*B*). Clades I and II both have sequences from all eukaryotic lineages including fungi (see fig. 7) whereas clades III–XI contain, besides two orphans, only fungal sequences. Clades III–XI occur at large distances from the LCA. Hence, APs have undergone a much more pronounced sequence diversification in fungi than in other eukaryotes. The high sequence diversification of fungal APs has been caused by either adaptation or drift. Sequences at relatively large distances (e.g., the sequences from clade XI, all from Ascomycetes and with a recent common ancestor) have likely been subject to drift. Another part of the sequence diversification appears specific to certain taxonomical groups, either as a result of mere drift, or a combination of drift and ecological adaptation. Clade IV consists of only Basidiomycete sequences, whereas clades VII and IX contain only Pezizomycete sequences (see fig. 7). Clade VIII contains only Ascomycete sequences and has Saccharomycete-specific and Pezizomycete-specific subclades (see fig. 5, discussed in more detail below). Most of the sequences of these clades have relatively short branches, which suggest constraint.

Gene duplication typically results in functional redundancy and allows for sequence diversification mediating potential functional specialization. Such functional diversification results in different functional constraints, which provoke different paralogs to evolve in different directions, and is reflected in the complex topology of the tree. The clearest evidence for functional diversification is the yapsin subfamily of clade III. Yapsins are well-characterized fungal APs that are involved in the proteolytic activation of peptide prohormones (Karlsen et al. 1998), or hydrolases, similar to the action of the serine protease Kexin (Schild et al. 2011). Indeed, the yapsins YPS1 and YPS2 from *S. cerevisiae* were identified as suppressors of a Kexin null mutant (Werten and de Wolf 2005). The distinct biological functions of yapsins, as compared with for instance the vacuolar APs from clade I or the secreted Aspergillopepsin from clade III, corresponds with their cellular location. Maturases need to operate within the secretory organelles, just as Kexin does (Jalving et al. 2000), or in the extracellular matrix. Yapsins typically possess a GPI-anchor by which they can be attached to the cell membrane or the cell wall (Gagnon-Arsenault et al. 2006). Figure 5*A* shows that many sequences in clade VIII have a GPI-anchor motif. Furthermore, the yapsins appear to have a high substrate specificity. Although APs are considered to prefer hydrophobic substrates, yapsins have a preference for

mono- or dibasic cleavage sites (Schild et al. 2011). The CDP analysis provided a possible explanation for this specificity. Figure 6 illustrates that the yapsins almost invariantly have a solvent exposed Glu residue at CDP129, near the active site. The other APs contain mostly polar residues such as Ser, Thr or, in the case of clade III, an Asn. Site-directed mutagenesis studies on Glu129 would give experimental insight into whether this residue contributes to the preference of yapsins for mono- and dibasic cleavage sites.

Although the majority of Saccharomycete yapsins that lack a GPI-anchor motif are paraphyletic according to both our phylogeny (fig. 5*B*) and the phylogeny recently demonstrated by Monod et al. (2011), this is still consistent with a duplication of the yapsin gene in the ancestor of the Saccharomycetes. Subsequently, one of the duplicates presumably lost the GPI-anchor motif. APs in both groups are functional as most sampled Saccharomycete species have paralogs in both groups. The exact function of yapsins lacking a GPI-anchor remains to be determined although a number of *Candida albicans* homologs in this subclade have been shown to be involved in pathogenesis (reviewed by Albrecht et al. 2006). Subclade ii of clade VIII (fig. 5) also contains barrierpepsin which is involved in the cleavage of the peptide pheromone α-factor (Monod et al. 2011), which has also been shown to be a yapsin substrate (Azaryan et al. 1993). Barrierpepsin is a secreted, heavily glycosylated AP (Jars et al. 1995) which lacks a GPI anchor but does contain the insert between Cys45 and Cys52 mentioned earlier. Hence, barrierpepsin appears to be an even more recent homolog and its function and substrate specificity exemplify the functional redundancy and diversification processes that follow a gene duplication event. A more detailed description is described by Monod et al. (2011).

It is not clear whether the Pezizomycete APs in clade IX represent another example of functional diversification of yapsins. These peptidases are similar to the yapsins, and many contain GPI-anchor motifs whereas others do not. It is also unclear whether the presence of a GPI-anchor motif in fungal APs, present in APs in clades IV, V, VIII and IX, is monophyletic as GPI-anchor motifs are low complexity regions that cannot be reliably aligned.

The phytopathogen subclade of clade II (fig. 4*B*) provides another example of functional diversification. Although many phytopathogenic fungi lack a homolog in this subclade, the absence of a representative from a saprophytic fungus suggests that these APs are related to virulence. The sequences within this subclade show two characteristics that point to functional specialization. First, all members of this subclade lack the propeptide typically present in AP precursors or zymogens, which implies that these enzymes enter the endoplasmic reticulum (ER) in an active form, rather than as zymogens. In addition, all sequences within this subclade show substitutions of Asp-317 and Ser/Thr-218, which may serve to extend the pH range over which the enzymes are stable and active. Renin, the AP which is involved in the regulation of blood pressure, is active at pH values as high as neutrality and contains Ala residues at both positions 317 and 218. Mutation of Ala-317 to the more conventional Asp significantly lowered the pH at which renin remained active (Yamauchi et al. 1988). A similar role for Ala-218 has been postulated based on the occurrence of this residue at the equivalent position, adjacent to the catalytic Asp residue in A2 retropepsins (Ido et al. 1991). The occurrence of these two substitutions together with the absence of a propeptide in the phytopathogen subclade of APs might thus be the result of an adaptation to a less acidic environment. The exact function of these peptidases remains to be elucidated.

Coclustering of features is also found in a subclade of clade III comprising sequences that lack a signal peptide as well as disulphide bridges (see fig. 4*C*). As a signal peptide is essential for targeting proteins to the ER and disulphide bridges are typically generated within the ER, these subclade members either remain confined to the cytosol or are subject to an alternative secretion pathway. It remains to be demonstrated whether this is another example of functional diversification.

Toward a Redefinition of the Classification of A1 APs

Our data indicate that of the 11 selected clusters of APs only two conform (largely) to the MEROPS A1A and A1B subfamilies. MEROPS A1A contains 111 sequences listed as "holotypes," 71 of which were included in our phylogeny. Of the latter, 30 fell into clade I, whereas 41 (almost 60%) clustered into six other clades (II, III, V, VI, VIII, and IX). Clade I contains sequences from various eukaryotic orders as discussed above and correlates well with A1A. All sequences classified by MEROPS as A1B fall into clade II. The variation of sequences in clades III–XI provides a stark contrast with the conservation of the sequences in clades I and II. Certain clades consist of APs with clearly distinguished functions and/or sequence characteristics related to functional aspects. The recent review on fungal APs (Monod et al. 2011) also clearly summarizes a number of functionally different fungal APs. Therefore, we propose a more comprehensive classification of A1 APs based on phylogenetic clustering. Clustering should be understood as grouping objects such that the objects in the same group or cluster are more similar to each other than to other objects. It should be performed using a comprehensive tree based on homologous characters only, that is, omitting cluster-specific subsequences, such as C-terminal extensions and the nepenthesin-specific insert. Classification should be understood as the assignment of a query sequence to a defined cluster and can easily include cluster-specific subsequences, thereby exploiting the complete information of the query sequence and the various clusters. Such a fast and reliable biocomputational classification of sequences will be required in order to efficiently characterize the increasing amount

of genomic data. Thus, based on the fungal AP diversification it seems appropriate to amend the current MEROPS classification of A1 APs based on dedicated phylogenetic clustering.

Supplementary Material

Supplementary datafiles S1–S3, tables S1 and S2, and figure S1 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Dr Joost Stassen for running the bootstrap analysis of the main phylogenetic tree. This work financially was supported by the Universidad Nacional de Mar del Plata (AtH) and the Consejo Nacional de Investigaciones Científicas y Técnicas (M.V.R., AtH).

Literature Cited

Abad-Zapatero C, et al. 1996. Structure of a secreted aspartic protease from *C. albicans* complexed with a potent inhibitor: implications for the design of antifungal agents. Protein Sci. 5:640–652.

Albrecht A, et al. 2006. Glycosylphosphatidylinositol-anchored proteases of *Candida albicans* target proteins necessary for both cellular processes and host-pathogen interactions. J Biol Chem. 281:688–694.

Alby K, Schaefer D, Bennett RJ. 2009. Homothallic and heterothallic mating in the opportunistic pathogen *Candida albicans*. Nature 460: 890–893.

Amselem J, et al. 2011. Genomic analysis of the necrotrophic fungal pathogens *Sclerotinia sclerotiorum* and *Botrytis cinerea*. PLoS Genet. 7: e1002230.

Anisimova M, Gascuel O. 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst Biol. 55: 539–552.

Anisimova M, Gil M, Dufayard J-F, Dessimoz C, Gascuel O. 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Syst Biol. 60: 685–699.

Athauda SBP, et al. 2004. Enzymic and structural characterization of nepenthesin, a unique member of a novel subfamily of aspartic proteinases. Biochem J. 381:295–306.

Azaryan AV, et al. 1993. Purification and characterization of a paired basic residue-specific yeast aspartic protease encoded by the YAP3 gene. Similarity to the mammalian pro-opiomelanocortin-converting enzyme. J Biol Chem. 268:11968–11975.

Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S. 2004. Feature-based prediction of non-classical and leaderless protein secretion. Protein Eng Des Sel. 17:349–356.

Berman HM, et al. 2000. The Protein Data Bank. Nucleic Acids Res. 28: 235–242.

Bondino HG, Valle EM, Ten Have A. 2012. Evolution and functional diversification of the small heat shock protein/α-crystallin family in higher plants. Planta 235:1299–1313.

Capra JA, Singh M. 2008. Characterization and prediction of residues determining protein functional specificity. Bioinformatics 24: 1473–1480.

Castro LFC, Lopes-Marques M, Gonçalves O, Wilson JM. 2012. The evolution of pepsinogen C genes in vertebrates: duplication, loss and functional diversification. PLoS One 7:e32852.

Cho SW, Kim N, Choi MU, Shin W. 2001. Structure of aspergillopepsin I from *Aspergillus phoenicis*: variations of the S1′-S2 subsite in aspartic proteinases. Acta Crystallogr D Biol Crystallogr. 57:948–956.

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.

Criscuolo A, Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol. 10:210.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome Res. 14:1188–1190.

De Juan D, Pazos F, Valencia A. 2013. Emerging methods in protein co-evolution. Nat Rev Genet. 14:249–261.

Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS Comput Biol. 7: e1002195.

Eirín-López JM, Rebordinos L, Rooney AP, Rozas J. 2012. The birth-and-death evolution of multigene families revisited. Genome Dyn. 7: 170–196.

Eisenhaber F, et al. 2003. Prediction of lipid posttranslational modifications and localization signals from protein sequences: big-Pi, NMT and PTS1. Nucleic Acids Res. 31:3631–3634.

Espino JJ, et al. 2010. The *Botrytis cinerea* early secretome. Proteomics 10: 3020–3034.

Fernández-Acero FJ, et al. 2010. 2-DE proteomic approach to the *Botrytis cinerea* secretome induced with different carbon sources and plant-based elicitors. Proteomics 10:2270–2280.

Fraser ME, Strynadka NC, Bartlett PA, Hanson JE, James MN. 1992. Crystallographic analysis of transition-state mimics bound to penicillopepsin: phosphorus-containing peptide analogues. Biochemistry 31: 5201–5214.

Gagnon-Arsenault I, Tremblay J, Bourbonnais Y. 2006. Fungal yapsins and cell wall: a unique family of aspartic peptidases for a distinctive cellular function. FEMS Yeast Res. 6:966–978.

Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307–321.

Haubrock M, Waack S, Gültas M, Tüysüz N. 2012. Coupled Mutation Finder: a new entropy-based method quantifying phylogenetic noise for the detection of compensatory mutations. BMC Bioinformatics 13: 225.

Horton P, et al. 2007. WoLF PSORT: protein localization predictor. Nucleic Acids Res. 35:W585–W587.

Huang Y, Niu B, Ying G, Fu L, Li W. 2010. CD-HIT: a web server for clustering and comparing biological sequences. Bioinformatics 26:5.

Huson DH, Scornavacca C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol. 61:1061–1067.

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.

Jalving R, van de Vondervoort PJ, Visser J, Schaap PJ. 2000. Characterization of the kexin-like maturase of *Aspergillus niger*. Appl Environ Microbiol. 66:363–368.

Jars MU, Osborn S, Forstrom J, MacKay VL. 1995. N- and O-glycosylation and phosphorylation of the bar secretion leader derived from the barrier protease of *Saccharomyces cerevisiae*. J Biol Chem. 270: 24810–24817.

Karlsen S, Hough E, Olsen RL. 1998. Structure and proposed amino-acid sequence of a pepsin from atlantic cod (*Gadus morhua*). Acta Crystallogr D Biol Crystallogr. 54:32–46.

Kay J, Meijer HJ, Ten Have A, Van Kan JAL. 2011. The aspartic proteinase family of three Phytophthora species. BMC Genomics 12:254.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 305:567–580.
Lemberg MK, Freeman M. 2007. Functional and evolutionary implications of enhanced genomic analysis of rhomboid intramembrane proteases. Genome Res. 17:1634–1646.
Letunic I, Bork P. 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. 39: W475–W478.
Li B, Wang W, Zong Y, Qin G, Tian S. 2012. Exploring pathogenic mechanisms of *Botrytis cinerea* secretome under different ambient pH based on comparative proteomic analysis. J Proteome Res. 11: 4249–4260.
Ma L-J, et al. 2009. Genomic analysis of the basal lineage fungus *Rhizopus oryzae* reveals a whole-genome duplication. PLoS Genet. 5:e1000549.
Mazin PV, et al. 2010. An automated stochastic approach to the identification of the protein specificity determinants and functional subfamilies. Algorithms Mol Biol. 5:29.
Monod M, et al. 2002. Secreted proteases from pathogenic fungi. Int J Med Microbiol. 292:405–419.
Monod M, Staib P, Reichard U, Jousson O. 2011. Fungal aspartic proteases as possible therapeutic targets. In: Ghosh AK, editor. Aspartic acid proteases as therapeutic targets. Weinheim (Germany): Wiley-VCh Verlag GmbH. p. 573–606.
Mukaiyama H, Iwaki T, Idiris A, Takegawa K. 2011. Processing and maturation of carboxypeptidase Y and alkaline phosphatase in *Schizosaccharomyces pombe*. Appl Microbiol Biotechnol. 90: 203–213.
Newman M, et al. 1993. X-ray analyses of aspartic proteinases. V. Structure and refinement at 2.0 A resolution of the aspartic proteinase from *Mucor pusillus*. J Mol Biol. 230:260–283.
Parr CL, Keates RAB, Bryksa BC, Ogawa M, Yada RY. 2007. The structure and function of *Saccharomyces cerevisiae* proteinase A. Yeast 24: 467–480.
Pei J, Kim B-H, Grishin NV. 2008. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 36: 2295–2300.
Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 8:785–786.
Rawlings ND, Barrett AJ, Bateman A. 2012. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 40:D343–D350.
Runeberg-Roos P, Törmäkangas K, Ostman A. 1991. Primary structure of a barley-grain aspartic proteinase. A plant aspartic proteinase resembling mammalian cathepsin D. Eur J Biochem. 202:1021–1027.
Schild L, et al. 2011. Proteolytic cleavage of covalently linked cell wall proteins by *Candida albicans* Sap9 and Sap10. Eukaryot Cell. 10: 98–109.
Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18:6097–6100.
Shah P, et al. 2009. Comparative proteomic analysis of *Botrytis cinerea* secretome. J Proteome Res. 8:1123–1130.
Takahashi K, et al. 2005. Nepenthesin, a unique member of a novel subfamily of aspartic proteinases: enzymatic and structural characteristics. Curr Protein Pept Sci. 6:513–525.
Tang J. 1979. Evolution in the structure and function of carboxyl proteases. Mol Cell Biochem. 26:93–109.
Ten Have A, et al. 2010. The *Botrytis cinerea* aspartic proteinase family. Fungal Genet Biol. 47:53–65.
Ten Have A, Dekkers E, Kay J, Phylip LH, Van Kan JAL. 2004. An aspartic proteinase gene family in the filamentous fungus *Botrytis cinerea* contains members with novel features. Microbiology 150: 2475–2489.
Werten MWT, de Wolf FA. 2005. Reduced proteolysis of secreted gelatin and Yps1-mediated alpha-factor leader processing in a Pichia pastoris kex2 disruptant. Appl Environ Microbiol. 71:2310–2317.
Xiao G, et al. 2011. Genome sequence of the insect pathogenic fungus *Cordyceps militaris*, a valued traditional chinese medicine. Genome Biol. 12:R116.
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.

Associate editor: Jay Storz