Independent Subtilases Expansions in Fungi Associated with Animals

Anna Muszewska,*1 John W. Taylor,2 Pawel Szczesny,1,3 and Marcin Grynberg1

1Institute of Biochemistry and Biophysics, Warsaw, Poland
2Department of Plant and Microbial Biology, University of California
3Department of Biology, Institute of Molecular Plant Biology, University of Warsaw, Warsaw, Poland

*Corresponding author: E-mail: musze@ibb.waw.pl.

Associate editor: Charles Delwiche

Abstract

Many socially important fungi encode an elevated number of subtilisin-like serine proteases, which have been shown to be involved in fungal mutualisms with grasses and in parasitism of insects, nematodes, plants, other fungi, and mammalian skin. These proteins have endopeptidase activities and constitute a significant part of fungal secretomes. Here, we use comparative genomics to investigate the relationship between the quality and quantity of serine proteases and the ability of fungi to cause disease in invertebrate and vertebrate animals. Our screen of previously unexamined fungi allowed us to annotate and identify nearly 1000 subtilisin-containing proteins and to describe six new categories of serine proteases. Architectures of predicted proteases reveal novel combinations of subtilisin domains with other, co-occurring domains. Phylogenetic analysis of the most common clade of fungal proteases, proteinase K, showed that gene family size changed independently in fungi, pathogenic to invertebrates (Hypocreales) and vertebrates (Onygenales). Interestingly, simultaneous expansions in the S8 and S53 families of subtilases in a single fungal species are rare. Our analysis finds that closely related systemic human pathogens may not show the same gene family expansions, and that related pathogens and nonpathogens may show the same type of gene family expansion. Therefore, the number of proteases does not appear to relate to pathogenicity. Instead, we hypothesize that the number of fungal serine proteases in a species is related to the use of the animal as a food source, whether it is dead or alive.

Key words: subtilases, fungi, systemic human pathogens, serine proteases.

Introduction

Subtilases are serine endopeptidases and are considered to be among the broad spectrum of degrading enzymes found in almost all organisms. Most subtilases are secreted and especially so in saprobi fungal fungi, where subtilases often constitute a dominant component of the secretome (Hu and St. Leger 2004). In symbiotic fungi, subtilisin-like secreted serine proteases have been shown to play an important role in both pathogenic (Sreedhar et al. 1999; Donatti et al. 2008; Fang et al. 2009) and mutualistic associations (Reddy et al. 1996; Bryant et al. 2009). The first to be reported were the cuticle-degrading proteases from entomopathogenic fungi (Donatti et al. 2008; Fang et al. 2009) and then proteases from nematophagous (Yang et al. 2005; Wang et al. 2006), mycoparasitic (Yan and Qian 2009), and plant pathogenic species (Reddy et al. 1996). These studies were followed by reports of subtilases from endosymbionts of grass (Reddy et al. 1996) and from dermatophytes (Monod 2008). Diverse evolutionary fungal lineages rely on subtilases as the key proteases involved in infection, for example, the insect pathogen, Metarhizium anisopliae (Bagga et al. 2004) and the human dermatophyte Trichophyton rubrum (Jousson et al. 2004).

There have been multiple attempts to classify the serine proteases, all of them designed before the availability of diverse sequenced fungal genomes. As a result, there is significant disorder in the classification. In this work, to classify fungal serine proteases, we began with the MEROPS (Rawlings et al. 2008) and SCOP database classifications (Andreeva et al. 2004) together with families of the superfam­ily of subtilisin-like proteases defined by Siezen et al. (2007).

Proteolytic enzymes are classified into families and clans on the basis of amino acid sequence similarity and catalytic mechanism. Serine peptidases of the clan SB (subtilases), according to the MEROPS peptidase classification, are divided into two families S8 (subtilisin-like proteases) and S53 (serine-carboxyl proteinases) as shown in figure 1.

The S8 family proteases, characterized by an Asp-His-Ser catalytic triad (DHS triad), are often accompanied, on either side, by other domains. A similar His-Asp-Ser catalytic triad is present in S1 protease family, what is described as a clear example of convergent evolution (Hedstrom 2002). Subtilases are widely used in industry as detergent enzymes (Gupta et al. 2002), as well as in laboratories (proteinase K, subtilisin in washing buffer). S8 proteases are divided into two subfamilies S8A and S8B. Most known S8 representatives are grouped in the subtilisin S8A subfamily, among them: Tritirachium album proteinase K, Aspergillus flavus oryzin, streptococcal Cs a peptidase, Aspergillus alkaline peptidase, Beauveria cuticle-degrading peptidase, and many more. Proteinase K, the key S8A protease representative, is one of the best-described biological molecules (Gunkel and Gassen...
Kexin and furin are the canonical S8B members (known as kexins). Several protein structures are known for S8 proteases, including human proprotein convertases, which are associated with cholesterol metabolism and are involved in multiple neurodegenerative disorders (Nakayama 1997). S53 serine-carboxyl proteinases include Pseudomonas sedolisin, Bacillus kumamolisin, Aspergillus oryzae aorsin, and human tripeptidyl peptidase. S53 proteins have a conserved Ser-Glu-Asp triad and usually have a propeptide (Siezen et al. 2007).

Our analysis of the abundant and newly available fungal genomic sequence began with re-annotation of the proteomes and rapidly showed the presence of previously undescribed subtilisin groups as well as novel combinations of S8 or S53 domains with nonprotease domains. The broad sampling of fungal genomes allowed us to search for correlations between fungal genome content and their lifestyles. When we focused on protease families that are associated with animal pathogenesis and that have significantly expanded, we discovered that the expansion of subtilases appears to be a convergent adaptation to animal hosts, once in Onygenales (fungi parasitic on mammals) and again in Clavicipitaceae (fungi parasitic on insects).

**Materials and Methods**

**Sequence Database Searches**

Sequences of known S8 proteases subtilisin (GI:46193755), kexin (GI:19115747), and proteinase K (GI:131077) were used as seeds in PSI-BLAST searches of the fungal subset of the non-redundant (nr) database (Wheeler et al. 2008). For S53 analysis, tripeptidyl peptidase SED3 (GI:146323370) was selected as a seed.

For each sequence, the search was carried out with expectation (e) value threshold $10^{-3}$ until no new sequences were found. Most diverse hits were used as seeds for next searches. When expectation (e) value threshold was set to $10^{-5}$, proteins with S53 domain were retrieved in further iterations (from 8 to 12). The profiles from Pfam (Finn et al. 2010), Interpro (Hunter et al. 2009), or SMART (Letunic et al. 2009) describe S8 and S53 with a common profile. Duplicated hits as well as uncompleted sequences were discarded. Only full-length sequences from Eurotiomycetidae were aligned together with MAFFT (Katoh et al. 2005) using the local alignment option. Sequences encoding incomplete catalytic triad were excluded from further analysis.

**Sequence Clustering**

To elucidate the relationships between and within subfamilies of the S8 clan (S8 and S53), CLANS was used (Frickey and Lupas 2004). CLANS is a Java-based utility which visualizes pair-wise sequence similarities. Proteins in the graph are represented as vertices and all-against-all BLAST high-scoring segment pairs as edges.

**Domain Architecture**

The domain architectures of all analyzed subgroups were predicted using InterproScan (Zdobnov and Apweiler 2001), CD-Search (Marchler-Bauer and Bryant 2004), SMART (Letunic et al. 2009), and HHpred (Soding et al. 2005). Hypothetical signal peptides were detected with SignalP (Emanuelsson et al. 2007). Many previously unreported topologies were detected. However, many of the discovered topologies have no support in expressed sequence tag data and may be a consequence of erroneous in silico translation.

**Phylogenetic Analysis**

Conserved columns from the multiple sequence alignment (Supplementary fig. 5, Supplementary Material online) have been chosen with TrimAl using the “strict” parameter set (Capella-Gutierrez et al. 2009). The best model for phylogenetic analysis was selected with ProtTest (Abascal et al. 2005) ProtTest consistently selected the LG+G+I (Le and Gascuel 2008) as the most suitable model and WAG+G+I (Whelan and Goldman 2001) as the second best model.

Maximum Likelihood (ML) analysis has been calculated on a PhyML (Guindon et al. 2009) on-line server at Montpellier using the ProtTest recommended model and ten random starting trees. Bayesian analyses were carried out in MrBayes (Ronquist and Huelsenbeck 2003) with the following settings: number of generations 1,000,000, WAG amino acid substitution model with gamma parameter and a proportion of invariable sites. WAG was the second best model according to ProtTest, and MrBayes has not implemented the LG model jet. Trees were visualized and colored in iTol (Letunic and Bork 2007).

**Results**

**The Data Set**

In order to elucidate the role of subtilases in fungi, we first carried out simple sequence searches. Five different starting
points were used to collect a representative data set of fungal S8/S53 proteases. These sequences were merged together with those in the new Pfam 24.0 database, using the PF00082 definition of the subtilase domain, which includes both the S53 and S8 domains (Siezen et al. 2007). PSI-BLAST searches revealed that the S53 and S8 domains are distinct; no member of the hit list of either category (both with S8 and S53 queries) could enter the hit list of the other when the threshold e-value was 0.001. This result is congruent with CLANS clustering, which showed that these two groups are easily separable (fig. 2). Most of the identified predicted proteins have well-conserved active sites and likely are functional. Many genomes encode multiple secreted proteases. The pattern of enrichment of the number of encoded proteases in a single species shows an inverse relationship between the number of S8 and S53 proteases. It is often observed that an elevated number of S8 proteases is accompanied by a lower number of S53 proteases, and conversely, genomes rich in S53 proteases are poor in S8 enzymes (supplementary fig. 3, Supplementary Material online). This situation is present in the Aspergillus niger genomes; for example, A. niger CBS 513.88 encodes nine S8 proteases and seven S53 proteases. A. niger ATCC 1015 has, respectively, 11 and 4. The mean S8 and S53 content in the fungal subset of NR database is about the same for both protease types. The highest representation of S8 in analyzed genomes, S8 S8 domains, was identified in an early-diverging ascomycete, Pneumocystis jiroveci (see supplementary fig. 4, Supplementary Material online). Although our analysis includes many new fungal genomes, we offer the caveat that not all proteins in the NR database are from fully sequenced genomes.

Sequence Clustering—New Groups
To elucidate the relationships within the subtilase (S8) clade, we conducted a clustering analysis. Structure similarity, as noted by Wlodawer et al. (2003), and common profiles in databases are indicators of close relationships between subfamilies. As alluded to above, when S8 and S53 domain similarities were analyzed using the CLANS program, the S53 clade was very compact and distant to all S8 representatives (fig. 2). Clustering analyses of whole proteins or of proteinase domains alone (without other protein domains) of both S8 and S53 resulted in identical distribution of particular sequences. CLANS clustering relies on sequence similarity so the clusters reflect the differences in the proteinase domain independently of differences in the protein architecture (i.e., domain composition and organization).

In contrast to the single compact clade of S53 proteases, S8 proteases are much more variable and constitute many subfamilies. Within the S8 clade, kexins (S8B) form the best-separated clade, which is distant to S8A proteins. This compact and well-separated kexin clade is equivalent to the MEROPS S8B subfamily. The S8A members have a more complex distribution in the clustering scheme, reflecting the subtilisin-like superfamily classification of Siezen et al. (2007). We found support for this classification in that our CLANS graph analysis (supplementary fig. 1, Supplementary Material online), which included bacterial representatives of all six previously reported S8A categories showed compact and well-separated clades for each category and the absence of subtilisins, thermitases, and lantibiotic peptidases in Fungi (Siezen et al. 2007).
However, we found that S8 proteases form more subgroups than previously described. Here, we identify six new S8 protease groups based on their amino acid sequence, in addition to known groups containing kexin (S8B), proteinase K, pyrolisin, and osf (oxidatively stable alkaline serine protease) (Saeki et al. 2000). To accommodate the unexpected diversity of the S8A subfamily, including the six new groups, we suggest a redefinition of the classification of the clades of S8A proteases. Table 1 summarizes the composition of the direct neighborhood of all three amino acids constituting the DHS catalytic triad in each S8 subfamily as well as the co-occurrences of additional protein domains. Species distribution and sequence number of new groups is very limited (Table 1). A detailed taxonomic distribution of all S8 clades is presented in supplementary fig. 4 (see Supplementary Material online). Most of the novel groups are present in only a few species classified to Pezizomycotina (synonym of Euscomycota; Spatafora et al. 2006). The central super-clade is composed of more than 600 protease K genes and is the most variable category analyzed. The proteinases of interest, that is, those known to be or suspected to be involved in pathogenesis and symbiosis fall into the protease K clade. Many of the proteases from animal-infecting fungi are localized in the central part of the super-clade, in a very dense and compact area. To examine relationships in a manner different from clustering, we subjected the protein K sequences belonging to animal-related fungi to a phylogenetic analysis, which we report below in the section “Phylogenetic Analyses”.

Domain Architectures: Subtilase and Propeptide Domains

All proteins found in sequence searches to contain subtilase domains were also subjected to domain architecture analysis. The common architecture of most of the analyzed proteases includes a propeptide and the enzymatic domain. Most of the sequences grouped together in the protease K clade possess an N-terminal subtilisin propeptide (Subtilisin_N/Inhibitor_I9, Pfam:PF05922), the cleavage of which activates the enzyme.

In addition to the common subtilase and propeptide domains, our analysis predicts new combinations of subtilase domains with other domains, which are presented below. All typical and atypical architectures are shown in figure 3. In both S8 and S53 family members, new domain combinations have been noted. In the newly discovered S8 groups, domain fusions have been found for groups 1–3 but not for groups 4–6 (fig. 3). In addition to the propeptide domains in S8, the other domain combinations include PA, DUF1034, and sugar hydrolyzing domains.

P450 Domains

Deng et al. (2007) suggested that P450 domain is related to lifestyle and exhibits high variation in genome localization and amino acid sequence. We have found the co-occurrence of proteinase K and P450 domains in one Magnaporthe grisea protein (MGG_12799, GI:145608536), which is the first example of such a domain architecture.

PA Domain

S8B proteases (kexins) share a common domain architecture; they usually have a single transmembrane motif and a proprotein C-terminal convertase domain (P_protein) (Pfam:PF01483).

Pyrolisins and osf proteases usually have a proteinase-associated (PA, Pfam:PF02225) domain (Mahon and Bateman 2000), which is found as an insertion in many other proteases, for example, A22B, M36, M28, and trypsin. The function of this domain remains unclear, although Luo and Hofmann (2001) suggested that it may be involved in recognition of the protein by vacuolar sorting mechanisms. The PA domain often co-occurs with the DUF1034 (Domain of Unknown Function 1034, Pfam:PF06280). DUF1034 has been described as a domain often identified in bacterial and plant proteins.

Sugar Hydrolyzing Domains

New group 2 members fuse with different sugar-hydrolyzing domains such as alpha-1,3-glucanase (glyco_hydro_71, Pfam:PF03659), chitinase class II group (glyco_hydro_18, PF00704), or pectin lyase (SCOP:51133). These carbohydrate-degrading domains are known to play a role in fungal pathogenicity (Ait-Lahsen et al. 2001; Yakoby et al. 2001). An artificial construct of a bifunctional protein with both protease and chitinase activities showed enhanced effect on insects’ cuticle (Fang et al. 2009). This type of domain architecture was found in animal-related fungi (Histoplasma capsulatum Gt: 225554237), in plant pathogenic fungi (Nectria haematococca Gt: 256728098), and in non-pathogenic organisms (Podospora anserina Gt: 170942241).

Repeats

In S8, we found new domain combinations with repeat sequences, such as, Ankyrin, WD40, PT repeats, and one example of a fusion with a cyclin domain (new groups 1 and 3). Repeat sequences are supposed to play a role in protein–protein interactions (Al-Khodor et al. 2010). Cyclins are famous for their role in cell cycle regulation (Ágáu and Fajas 2010). Unexpectedly, an Aspergillus terreus protein (ATEG_02636, GI: 115388617) has two cyclin domains (InterPro: IPR006670) in the C-terminal location to the enzymatic domain. The fused protein may have a modified way of functioning and gain new regulatory abilities. Members of the new groups 4, 5, and 6 possess the protease domains only.

Most of the analyzed S53 proteases display the canonical catalytic domain and an inactivating propeptide architecture found in both kexins and proteinases K. Cleavage of an alpha and beta sandwich folded propeptide (Pro-kuma_activ, Pfam:PF09286) is necessary to activate the enzyme.

In S53 proteases, we noted some new domain co-occurrences. These include Sir2 and SAC3/GANP/Nin1/mts3/elf-3 p25 families. Sir2—sirtuin domain (Pfam:PF02146, silent information regulator 2) is one of the most intensively studied NAD-dependent protein deacetylases (North and Verdin 2004). A co-occurrence of Sir2 and S53 protease domains
Table 1. Active site and domain co-occurrence variability among S8 and S53 proteases. Columns DTG, GHGTS, and SGTS represent the closest amino acid sequence for each of the amino acids from the DHS catalytic triad.

<table>
<thead>
<tr>
<th>CLANS clade</th>
<th>DTG</th>
<th>GHGTH</th>
<th>SGTS</th>
<th>Additional domains</th>
<th>Taxonomic distribution</th>
<th>Number of sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>S8: new1</td>
<td>EP[VI][KR][IV][A][IV][LI]</td>
<td>HGT[HF][VI][AGL][LV][LK][VL]</td>
<td>SGTS[VF][AS][TPIA][AV][GA][IL]</td>
<td>[AV][AN]</td>
<td>WD40, ANK, TM, p-loop containing (PF00004 or PF05729)</td>
<td>Pezizomycotina 36</td>
</tr>
<tr>
<td>S8: new2</td>
<td>[EK][D][LF][GK][V][DG][EQ][FLIATEH]</td>
<td>[D][YW][P][V][G][A][L][Q][P][D][N][VL][EK]</td>
<td>[SATLV][SADH][GTS][LY][AS][PA]</td>
<td>[FL][VL][LA][G][SA][G][LV]</td>
<td>glyco_hydro_71 (PF03659) glyco_hydro_18 (PF00704), pectin lyase-like superfamily ANK, cyclin (IPR006670)</td>
<td>Pezizomycotina 23</td>
</tr>
<tr>
<td>S8: new3</td>
<td>[PR][VI][KVA][LI][DDG][VI]D</td>
<td>[P][YW][YW][SAxGHTIMA]</td>
<td>K[PS][VxY][TS][G][V][AS]</td>
<td>TALAAGL[AS][IV][LY][CVR]</td>
<td>glyco_hydro_71 (PF03659) glyco_hydro_18 (PF00704), pectin lyase-like superfamily ANK, cyclin (IPR006670)</td>
<td>Pezizomycotina 39</td>
</tr>
<tr>
<td>S8: new4</td>
<td>[KR][FY][P][DE][FE][GR][GN][V][RTV][V][AC][IV][L][DTG][V][DP][AG][A][AILP][GL]</td>
<td>HGTHVAG[IV][GA][HNQR][HT][PDQ][E][DP][HPQ][YNFW][YPFP][ADNR][HT][AS]</td>
<td>LG[DN][GES][KN][ST][M]</td>
<td>[DV][IV][TY][DG][SD][GTS][GA][SA]</td>
<td>—</td>
<td>Pezizomycotina and Taphrinomycotina 8</td>
</tr>
<tr>
<td>S8: new5</td>
<td>GINA[RL][YA][WT][CT][FI][PT]</td>
<td>[G][DG][AL][G][R][N][TV][N][GT][N][I][D]</td>
<td>[V][LT][G][EA][IQ][ML][LF][MG][Q][VAD][DV]</td>
<td>[SQ][Q][VE][GTS][VLE][AV][TW]</td>
<td>—</td>
<td>Onygenales 5</td>
</tr>
<tr>
<td>S8: new6</td>
<td>DIP[AV][I][YF][VDTG][AQ][IL][DN][NP][Q]</td>
<td>[T][E][Y][QT][E][G][G][Q][V][AC][VA][AC][AG]</td>
<td>[DP][DG][GH][TH][CA][G][S][V][I]</td>
<td>[DP][Q][W][Y][MF][F][IL][AS][GTS]</td>
<td>—</td>
<td>Pezizomycotina 17</td>
</tr>
<tr>
<td>S8: osf</td>
<td>TEY[QT][EG][G][Q][V][AC][VA][AC][AG]</td>
<td>DTGFD</td>
<td>[K][G][K][SD][T][DT][D]</td>
<td>[CG][G][H][TH][V][AS][G][I][A][AG]</td>
<td>—</td>
<td>Pezizomycotina and Taphrinomycotina 8</td>
</tr>
<tr>
<td>S8: pyrolisin</td>
<td>VKDL[HR][A][EQ][G][TL][JK][G][V][K][VI][V][AG]</td>
<td>[AG]</td>
<td>[V][G][I][IV][D][T][G][I][D]</td>
<td>[V][DY][TH][PA][LG]</td>
<td>—</td>
<td>Basidiomycota, Pezizomycotina and Pichia 75</td>
</tr>
<tr>
<td>S8: proteinase K</td>
<td>[I][VL][D][TS][G][I][V][RN][IV][THP][ED]</td>
<td>DNG[GH][HT][V][C][A][T][I][J][G][A]</td>
<td>[G][S][T][M][AG][V][AK][N][K][N][L][VI][AV][K]</td>
<td>[AV][K][V]</td>
<td>—</td>
<td>Pezizomycotina and Taphrinomycotina 8</td>
</tr>
<tr>
<td>S8: proteinase K</td>
<td>[F][E][G][GRA]</td>
<td>[G][S][K][T][V][Y][G][A][K][N][K][N][L][VI][AV][K]</td>
<td>[AV][K][V]</td>
<td>[G][S][T][M][AG][V][AK][N][K][N][L][VI][AV][K]</td>
<td>—</td>
<td>Pezizomycotina and Taphrinomycotina 8</td>
</tr>
<tr>
<td>S8: pyrolisin</td>
<td>[V][D][G][G][D][L][Y][M][ET][SN][ED][K][AP]</td>
<td>[N][F][Y][N][NF][AEGS]</td>
<td>ExxxD</td>
<td>[V][DY][TH][PA][LG]</td>
<td>—</td>
<td>Pezizomycotina and Taphrinomycotina 8</td>
</tr>
<tr>
<td>S8: pyrolisin</td>
<td>[E][AGS][N][TD][L][D][IV][OE][Y][AI][V][G][L][I][L][S][P][V][Q][T][E][VL][Y][F][SQ]</td>
<td>[V][L][S][T][Y][V][E][R][ED][IN][EQ][V][L][P][S][Y][A][V][Q][C][N][E][L][E][F][Y][A][G][Q][K]</td>
<td>LG[AL][L][R][L][G][V][ST][L][J][L][F][S][A][GDS][S]</td>
<td>Gxx[x][L][V][G][T][G][S][A][S][T][P][Y][V]</td>
<td>Pro-kuma_activ (SM00944), Sir2 (PF001483), TM</td>
<td>Pezizomycotina and Pichia 190</td>
</tr>
</tbody>
</table>

Conserved motifs were predicted with the MEME tool (Bailey and Elkan 1994). Additional domains were detected with InterProScan, SMART, CD-search, SignalP, and HHpred. The number of sequences corresponding to each clade was directly obtained from a CLANS graph.

doi:10.1093/molbev/msr176
is found in the sequence of a *Gibberella zeae* hypothetical protein (Gl:46111169). Both cyclins and sirtuins are involved in cell cycle progression, therefore, the domain fusion may be crucial for a specific proteolysis that depends on the cell cycle point (Brachmann et al. 1995; McGowan 2003). Sirtuins are also known to affect the microtubule function, which may be important for motility (North and Verdin 2004). One may speculate the involvement of the Sir2-S53 fusion in the host attack.

The SAC3/GANP/Nin1/mts3/elf-3 p25 family domain (Pfam:PF03399) is found in various unrelated proteins. This domain is only defined by some structurally conserved loci and is known to appear in proteins belonging to big complexes. Possibly, the domain itself is important for protein interactions and plays a similar role for proteases (Kom immami and Toh-e 1994; Gordon et al. 1996; Seeger et al. 1996; Takei and Tsujimoto 1998; Jones et al. 2000; Kuwahara et al. 2000; Burks et al. 2001). An *A. terreus* protein (Gl:115384808) has a C-terminal domain similar to this domain, apart from its Pro-kuma_activ propeptide.

### Phylogenetic Analyses

Phylogenetic relationships were analyzed for a set of 103 sequences of Onygenales that included vertebrate pathogens and their non-pathogenic relatives. The phylogenetic analysis was conducted with two different methods: Bayesian analysis (BA) and ML. Trees obtained from both methods had the same topology as seen in figure 4. An additional analysis using the same approach was made for a set of 102 sequences of invertebrate pathogens from the Hypocreales together with Onygenales to verify whether the invertebrate and vertebrate animal pathogens share similar expansions (Supplementary fig. 2, Supplementary Material online). The tree was rooted with *T. album* protease K sequence (Gl:131077) and has a well-supported topology. Our observations indicate that *M. anisopliae* protease K (Hypocreales) expanded and diversified independently from those in Onygenales.

### The Fungi Pathogenic against Vertebrates (Onygenales)

The subtilisin-like serine proteases have been shown to play a major role in skin infection of mammals (Descamps et al. 2002). Now that fungi responsible for systemic disease have been sequenced, we had the opportunity to see whether the subtilase enrichment is a common feature of Onygenales or not. We analyzed a set of sequences from both systemic and cutaneous-related Onygenales (fig. 4). In this analysis, we will name subtilase clades in concordance with the *T. rubrum* subtilases nomenclature proposed by Monod and his collaborators, that is, as SUBx, where x is an ascending number (Jousson et al. 2004; Monod 2008). The traditional names of *A. niger* proteases (PepC and PepD) will be kept. We rooted our phylogenetic tree with the *T. album* protease K, and both the position of the root and the order of branching of the deepest divergences are not well supported (fig. 4). Clades named based on the presence of *Aspergillus* PepC and PepD as well as the group
consisting of PepC, PepD, and SUB2 are well supported (with a bootstrap value of 1.00), which is crucial for the reliability of the rest of the analysis. The PepD, PepD, and SUB2 clades possibly originate from a duplication event before the Eurotiales/Onygenales split. The SUBs of the dermatophytic Onygenales seem to have evolved by a series of duplication events after the split of the main lineages (fig. 4, each SUB is marked with a separate color).

The phylogenetic analysis of subtilisin-like serine proteases shows that Onygenales have representatives in all clades except for PepD, whereas Eurotiales are represented in just two clades, PepC and PepD. Because our data set is composed mostly of Onygenales sequences, we will concentrate on this order and not Eurotiales.

After the divergence of the Eurotiales and the Onygenales lineages, many duplications occurred leading to the dichotomous architecture of the tree. There are duplication events that happened before the divergences of the Ajellomycetaceae (Paracoccidioides and Histoplasma), Arthodermataceae (Microsporum and Trichophyton), and Onygenaceae (Coccidioides and Uncinocarpus) as documented by the presence of PepC, SUB6, 7, and 8 in all tree lineages (Arthodermataceae, Ajellomycetaceae, and Onygenaceae). Some duplications were retained both in Arthodermataceae and Onygenales, this is the case of SUBs 1, 3, 4, 5, 6, and 7 show duplication events that happened before the divergences.
events, but only among Arthrodermataceae (fig. 4). In addition, Onygenaceae-specific duplications can be observed in SUB2 and SUB12–17 clades. These proteases have been named with the following numbers continuing the Monod’s naming system (Jousson et al. 2004; Monod 2008).

All of the analyzed dermatophytic and systemic Onygenales share a highly similar number of encoded proteases K, suggesting an ancestral formation of the core protease set before the Arthrodermataceae and the Onygenales split. The Eurotiales have no SUB3, SUB4, SUB1, SUB5, SUB6, and SUB7 homologous sequences, so the common ancestor of Eurotiomycetes might have had a limited subtilisin repertoire compared with the wealth of SUBs in Onygenales. The alternative scenario with a subsequent loss of all SUBs in Eurotiales lineage seems less likely but not impossible. Duplication events followed by retention of both of the copies seem to be a common event in the evolutionary history of protease K sequences in Onygenales. Signs of successful duplications can be observed at different scales ranging from family specific (SUB3 vs. SUB4 and SUB1 vs. SUB5) to order-wide conserved (SUB6 vs. SUB7, SUB8 vs. SUB6 and SUB7). We show that dermatophytic fungi in the Arthrodermataceae share with members of the Coccidioides group, an elevated number of phylogenetically close protease K genes. It is tempting to think that dermatophytes and systemic fungal pathogens, for example, Coccidioides species, share the abundance of this type of subtilisin genes (Coccidioides spp and Uncinocarpus have 16 S8 protease genes), however, another group of systemic fungal pathogens, the Arthrodermataceae, does not encode an elevated number of protease K genes, for example, Paracoccidioides and Histoplasma have only six proteases (Sharpton et al. 2009). Phylogenetically, a duplication of protease K genes must have occurred before the divergence of the PepC plus PepD plus SUB2 clade from all the remaining clades, but following that ancestral duplication, subsequent duplications that occurred early in both of these clades must have been followed by the loss of the duplicated proteases in the Arthrodermataceae. As a result, although dermatophytes subtilisin-like serine proteases have orthologs in Coccidioides and Uncinocarpus, they do not have them in the Arthrodermataceae. Simply, being capable of systemic human infection does not imply an elevated number of protease K genes.

**Discussion**

Proteolytic enzymes are well known to be involved in host–pathogen interactions. The subtilisin family appears to play many roles in fungal biology. To get a complete view of fungal subtilases, we searched available protein sequence data to find all fungal subtilase domains. We found more than a thousand fungal subtilases in the NCBI protein NR database and submitted them to clustering analysis to develop a new classification of the S8 and S53 domains. Sequence clustering showed clearly that S8 and S53 constitute discrete categories. The S8A proteases comprise a variety of poorly defined and distantly related categories in contrast to the S8B proteases, which are very well defined and easily distinguished. Based on our clustering results and Siezen’s reviews (Siezen et al. 2007), we suggest a revision of subtilisin-like serine protease subfamilies that splits the S8A subfamily into smaller better-defined subgroups. We characterized six new groups of fungal subtilases, most of which, interestingly, have a limited taxonomic distribution suggesting a narrow specialization. These observations need further experimental study because, with bioinformatic tools alone, we cannot describe their biological and biochemical properties.

The clustering analysis not only showed new categories but also showed expansions of gene families. Our studies are consistent with previous information that some serine proteases expanded in filamentous fungi (Bagga et al. 2004). Although most fungi gain nutrition as symbionts or decomposers of plants, the fungal species associated with these expansions of proteins containing S8A domains are associated with animals. These fungi fall into two fungal clades, one associated with invertebrate animals, Clavicipitaceae in the Hypocreales, and the other with vertebrates, Onygenales. Both groups have been experimentally shown to use subtilases in animal infections (Descamps et al. 2002; Jousson et al. 2004). The phylogeny is consistent with independent expansions of these protease families in Clavicipitaceae and Onygenales.

Considering the finding that subtilases have been associated with pathogenesis against multiple hosts, we hypothesize that they may play a role in a common evolutionary strategy in fungi. Analyzing multiple genomes enabled us to observe correlations that have not been noted before. As presented in supplementary figs. 3 and 4 (see Supplementary Material online), the number of encoded subtilases for different evolutionary lineages is variable. The number of subtilases per genome cannot be a discriminative criterion in assuming the ecological niche. Protease gene family expansions appear to be an important evolutionary step among the fungi that show a long association with animals but not necessarily a sufficient step to define virulence because some systemic human pathogenic fungi show extreme expansions (e.g., Coccidioides or Pneumocystis jiroveci) and others do not (e.g., Histoplasma). These new protease K genes were named following Monod’s method for naming of Trichophyton SUB proteases (Jousson et al. 2004; Monod 2008). We expect that this classification may change as new sequencing data become available. The association between protease gene family expansions and pathogenicity does not extend to fungi that are opportunistic pathogens (e.g., Aspergillus fumigatus). Our data suggest that S8 proteases can be involved in infections not as a virulence factor per se, but by the use of animal protein, whether living or dead, as a primary substrate. There are fungi with expanded families that do not cause human disease, for example, Uncinocarpus reesei and others that do cause disease but lack the expanded families, for example, Histoplasma and Blastomyces. This phenomenon indicates that the expansion, itself, is not the key factor that distinguishes
pathogens from nonpathogens. Of course, the expanded families of proteases may be pathogenicity factors in the sense that their absence would render the fungus incapable of causing disease. However, proving subtilase function may be technically complicated because of the elevated number of subtilases in systemic infection fungi, for example, there are 16 proteins with subtilase domains in the *Coccidioides* genomes. Deleting one protease K gene may not lead to any interesting phenotype. In fact, taking into consideration the example of *M. anisopliae* (Bagga et al. 2004) and *T. rubrum* (Jousson et al. 2004), we expect to find multiple subtilases to be involved in infection rather than a single “pathogenic” protease. Neither do we know how the genes are regulated or whether they are co-regulated, which seems likely. Applying bioinformatic tools enabled us to analyze many proteomes at a time and to observe protein evolution at genomic level. The next step in the analysis of pathogenicity would be a thorough search and analysis of all proteases in fungal genomes.

It is very likely that the content of the secreted protease cocktail can be adapted in many ways in order to suit a specific ecological niche. The general inverse relationship between the number of encoded S53 and S8 proteases suggests some compensation mechanism. One possible explanation for the observation that all fungi have S8 serine proteases, whereas some lineages lack S53 serine proteases, is that S8 have a broader function and S53 are more specific. S53 serine proteases, although less studied in fungi, may play an important role in interactions with the environment and especially in plant pathogenic fungi.

Our interpretation of subtilisin evolution has emphasized gene duplications over gene losses. For example, the Arthrodertmatacea and Onygena share a large number of proteases K genes, whereas the Eurotiales have no examples from clades SUB3, SUB4, SUB1, SUB5, SUB6, and SUB7. Although it is formally possible that the duplications occurred before the divergence of Onygena and Eurotiales, and copies were then lost from the Eurotiales, we consider it more likely that the gene duplications occurred in the ancestor of the Onygena, after their divergence from the Eurotiales. In favor of our thinking is the observation that duplication events are commonly followed by retention of both of the copies in the evolutionary history of proteinase K sequences in Onygena. Signs of successful duplications can be observed at different scales ranging from family specific to order-wide conserved. However, there are undoubtedly cases where duplicated genes have been lost.

The evolutionary history of proteinase K sequences is a story of duplication events. Given that some of the analyzed organisms showed a strong tendency toward duplication retention (*Coccidioides, Microsporum*), whereas others were more conservative (*Ajellomycetaceae*), one wonders if it is the tendency to duplicate or the retention that explains the differences in gene family size between lineages. We favor the explanation that duplication events are similar in different lineages, but that selection for the retention of duplicated genes is the key event that drives the differences in gene family number on different lineages, as has been seen for segmental duplications in yeast (Dujon 2010).

### Supplementary Material

Supplementary figures S1–S5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

### Acknowledgments

The authors are thankful to Prof. Andrzej Paszewski for suggestions and assistance and to Prof. Gustavo Goldman for guidance and support. J.W.T. acknowledges support from NIH-NIAID 1R01AI070891 and NIH-US4AI65359.

### References


