MBE Advance Access published July 4, 2011

Independent Subtilases Expansions in Fungi Associated With Animals

Research Article

Anna Muszewska[1], John W. Taylor[2], Paweł Szczęsny[1,3], Marcin Grynberg[1]

[1] Institute of Biochemistry and Biophysics, Pawinskiego 5A, 02-106 Warsaw, Poland

[2] Plant and Microbial Biology, 111 Koshland Hall, University of California, Berkeley, CA 94720-3102, USA

[3] Institute of Molecular Plant Biology, Department of Biology, University of Warsaw, Miecznikowa 1, 02-096 Warsaw, Poland

Corresponding author: Anna Muszewska, address as above; musze@ibb.waw.pl, +4822 5921315 (tel); +4822 592 21 90 (fax)

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

Running head: Subtilase Expansion in Fungi

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.

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 saprobic fungi, where subtilases often constitute a dominate 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, Lam, and Belanger 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, Lam, and Belanger 1996). These studies were followed by reports of subtilases from endosymbionts of grass (Reddy, Lam, and Belanger 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 classifications (Rawlings et al. 2008) and SCOP (Andreeva et al. 2004) together with families of the superfamily of subtilisin-like proteases defined by Siezen and colleagues (Siezen, Renckens, and Boekhorst 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 proteinases) and S53 (serine-carboxyl proteinases) as shown in Figure1.

The S8 family proteases, characterized by an Asp-His-Ser catalytic 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, Beg, and Lorenz 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 C5a peptidase, *Aspergillus* alkaline peptidase, *Beauveria* cuticle-degrading peptidase and many more. Proteinase K, the key S8A proteinase representative, is one of the best described biological molecules (Gunkel and Gassen 1989). 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, Renckens, and Boekhorst 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).

Results

The dataset

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 dataset 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, Renckens, and Boekhorst 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 material Fig. 3). This situation is present in the *Aspergillus niger* genomes; for example, *A. niger* CBS 513.88 encodes 9 S8 proteases and 7 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, 58 S8 domains, was identified in an early-diverging ascomycete, *Pneumocystis jiroveci* (see Supplementary data Fig. 4). 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 (SB) clade we conducted a clustering analysis. Structure similarity, as noted by Wlodawer and colleagues (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. (Siezen, Renckens, and Boekhorst 2007). We found support for this classification in that our CLANS graph analysis (Supplementary material Fig. 1), 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, Renckens, and Boekhorst 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 proteins. 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 material (Fig. 4). Most of the novel groups are present only in a few species classified to Pezizomycotina (synonym of Euascomycota, 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 Fig. 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 hydrolysing domains.

P450 domains. Deng and colleagues suggested that P450 domain is related to lifestyle and exhibits high variation in genome localization and amino acid sequence (Deng, Carbone, and Dean 2007). 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, e.g. A22B, M36, M28, trypsin. The function of this domain remains unclear, although Luo and Hofmann (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 pathogenecity (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* GI: 225554237), in plant pathogenic fungi (*Nectria haematococca* GI: 256728098) and in non-pathogenic organisms (*Podospora anserina* GI: 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 (Aguilar 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 proteins 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/eIF-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-ocurrence of Sir2 and S53 protease domains is found in the sequence of a *Gibberella zeae* hypothetical protein (GI: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/eIF-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 (Kominami 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 *Aspergillus terreus* protein (GI:115384808), has a C-terminal domain similar to this domain, apart from its Pro-kuma_activ propeptide.

Phylogenetic analyses

Phylogenetic relationships were analysed 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 Maximum Likelihood (ML). Trees obtained from both methods had the same topology as seen in Fig. 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 Material Fig. 2). The tree was rooted with *Tritirachium album* protease K sequence (GI:131077) and has a well supported topology. Our observations indicate that *Metarhizium 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 analysed 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 *Trychophyton 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 *Aspergillus niger* proteases (PepC and PepD) will be kept. We rooted our phylogenetic tree with the *Tritirachium 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, 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 colour). 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 dataset 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*, *Histoplasma*), Arthodermataceae(*Microsporum*, *Trichophyton*) and Onygenaceae (*Coccidioides*, *Uncinocarpus*), as documented by the presence of PepC, SUB6,7,8 in all tree lineages (Arthordermataceae, Ajellomycetaceae, Onygenaceae). Some duplications were retained both in Arthrodermataceae and Onygenales, this is the case of SUB5 and SUB1 versus SUB9. There are some cases where duplications must have occurred with a subsequent loss of one copy, for example in SUB9 there is a duplication that must have occurred before the divergence of *Coccidioides* and *Uncinocarpus*, but *Uncinocarpus* retains only one of the duplicates. Other duplication events are specific to one lineage, for example SUBs 1, 3, 4, 5, 6, and 7 show duplication 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 analysed 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 Onygenaceae 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 to 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 seems to be a common event in the evolutionary history of proteinase K sequences in Onygenales. Signs of successful duplications can be observed at different scales ranging from family specific (SUB3 v. SUB4, SUB1 v. SUB5) to order-wide conserved (SUB6 v. SUB7, SUB8 v. SUB6&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, e.g. *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, e.g. *Paracoccidioides* and *Histoplasma* have only 6 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 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, Renckens, and Boekhorst 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). Whereas 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 goups 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 Figures 3 and 4, 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, e.g. *Uncinocarpus reesei*, and others that do cause disease but lack the expanded families, e.g. *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 e.g. 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 *Metarhizium anisopliae* (Bagga et al. 2004) and *Trichophyton 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 Arthrodermataceae and Onygenaceae 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 Onygenales 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 Onygenales, 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 Onygenales. 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 analysed organisms showed a strong tendency towards 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).

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 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 10-2, proteins with S53 domain were retrieved in further iterations (from 8-12). The profiles from Pfam (Finn et al. 2010), Interpro (Hunter et al. 2009) or SMART (Letunic, Doerks, and Bork 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 SB 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 (HSPs) 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, Doerks, and Bork 2009) and HHpred (Soding, Biegert, and Lupas 2005). Hypothetical signal peptides were detected with SignalP (Emanuelsson et al. 2007). Many previously unreported topologies where detected. However, many of the discovered topologies have no support in EST data and may be a consequence of erroneous *in silico* translation.

Phylogenetic analysis

Conserved columns from the MSA (Supplementary Material Fig. 5) have been chosen with TrimAl using the “strict” parameter set (Capella-Gutierrez, Silla-Martinez, and Gabaldon 2009). The best model for phylogenetic analysis was selected with ProtTest (Abascal, Zardoya, and Posada 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 analysis has been calculated on a PhyML (Guindon et al. 2009) on-line server at Montpellier using the ProtTest recommended model and 10 random starting trees. Bayesian analyses were carried out in MrBayes (Ronquist and Huelsenbeck 2003) with the following settings: number of generations 1000000, 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 coloured in iTol (Letunic and Bork 2007).

Supplementary Material

Supplementary Figure 1: 2D CLANS clustering of 1100 fungal S8 and S53 proteases (obtained from iterative BLAST searches against the fungal subset of the NR database and Pfam database) and 60 bacterial S8 proteases (10 representative sequences of each 6 subtilisin like serine proteases subfamily described by Siezen and colleagues (42)). The reference sequences are depicted as triangles.

Supplementary Figure 2: Phylogenetic tree of 104 proteinase K sequences. Maximum-likelihood analysis of a set of proteases was carried out using the LG +G model. Species abbreviations: Met. - *Metarhizium*, Asp. - *Aspergillus*, Epi. - *Epichloe*, Bea. - *Beauveria*, Tri.- *Trichophyton*, Mic. -

Microsporium, Coc. *Coccidioides*, Unc. *Uncinocarpus*.

Supplementary Figure 3: The variability of S53 and S8 families size in Fungi. Sequences were collected by BLAST searches against the fungal subset of the protein NR database and directly from the Pfam database. The schema prepaired in iTol (Letunic and Bork 2007) presents the number of encoded proteins with S53 and S8 protease domain per organism present in the NR database. Sequence count is based on the presence of taxid (taxonomic id from NCBI database). There are many unsequenced and incomplete genomes present in the dataset (protein NR database).

Supplementary Figure 4: 2D CLANS clustering of S8 proteases resulted in 10 distinguishable clades. The schema prepaired in iTol (Letunic and Bork 2007) shows the number of encoded proteins of each clade per organism present in the NR database. The colours used in the 2D clustering image are the same as on the pseudo-tree.

Supplementary Figure 5: Sequence alignment of Onygenales family proteinase K proteases. Columns highlighted in blue were selected by TrimAl and used for phylogenetic analysis. Sequence s are named using their GI number (GI is a unique identifier from NCBI protein database).

Acknowledgments: The authors are thankful to prof. Andrzej Paszewski for suggestions and assistance and to prof. Gustavo Goldman for guidance and support. JWT acknowledges support from NIH-NIAID 1R01AI070891 and NIH-U54-AI65359.

Literature cited

Abascal, F., R. Zardoya, and D. Posada. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics **21**:2104-2105.

Aguilar, V., and L. Fajas. 2010. Cycling through metabolism. EMBO Mol Med **2**:338-348.

Ait-Lahsen, H., A. Soler, M. Rey, J. de La Cruz, E. Monte, and A. Llobell. 2001. An antifungal exo-alpha-1,3-glucanase (AGN13.1) from the biocontrol fungus *Trichoderma harzianum*. Appl Environ Microbiol **67**:5833-5839.

Al-Khodor, S., C. T. Price, A. Kalia, and Y. Abu Kwaik. 2010. Functional diversity of ankyrin repeats in microbial proteins. Trends Microbiol **18**:132-139.

Andreeva, A., D. Howorth, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin. 2004. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res **32**:D226-229.

Bagga, S., G. Hu, S. E. Screen, and R. J. St. Leger. 2004. Reconstructing the diversification of subtilisins in the pathogenic fungus *Metarhizium anisopliae*. Gene **324**:159-169.

Brachmann, C. B., J. M. Sherman, S. E. Devine, E. E. Cameron, L. Pillus, and J. D. Boeke. 1995. The SIR2 gene family, conserved from bacteria to humans, functions in silencing, cell cycle progression, and chromosome stability. Genes Dev **9**:2888-2902.

Bryant, M. K., C. L. Schardl, U. Hesse, and B. Scott. 2009. Evolution of a subtilisin-like protease gene family in the grass endophytic fungus *Epichloe festucae*. BMC Evol Biol **9**:168.

Burks, E. A., P. P. Bezerra, H. Le, D. R. Gallie, and K. S. Browning. 2001. Plant initiation factor 3 subunit composition resembles mammalian initiation factor 3 and has a novel subunit. J Biol Chem **276**:2122-2131.

Capella-Gutierrez, S., J. M. Silla-Martinez, and T. Gabaldon. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics **25**:1972-1973.

Deng, J., I. Carbone, and R. A. Dean. 2007. The evolutionary history of cytochrome P450 genes in four filamentous Ascomycetes. BMC Evol Biol **7**:30.

Descamps, F., F. Brouta, M. Monod, C. Zaugg, D. Baar, B. Losson, and B. Mignon. 2002. Isolation of a *Microsporum canis* gene family encoding three subtilisin-like proteases expressed in vivo. J Invest Dermatol **119**:830-835.

Donatti, A. C., L. Furlaneto-Maia, M. H. Fungaro, and M. C. Furlaneto. 2008. Production and regulation of cuticle-degrading proteases from *Beauveria bassiana* in the presence of *Rhammatocerus schistocercoides* cuticle. Curr Microbiol **56**:256-260.

Dujon, B. 2010. Yeast evolutionary genomics. Nat Rev Genet **11**:512-524.

Emanuelsson, O., S. Brunak, G. von Heijne, and H. Nielsen. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc **2**:953-971.

Fang, W., J. Feng, Y. Fan, Y. Zhang, M. J. Bidochka, R. J. St. Leger, and Y. Pei. 2009. Expressing a fusion protein with protease and chitinase activities increases the virulence of the insect pathogen *Beauveria bassiana*. J Invertebr Pathol **102**:155-159.

Finn, R. D., J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. Sonnhammer, S. R. Eddy, and A. Bateman. 2010. The Pfam protein families database. Nucleic Acids Res **38**:D211-222.

Frickey, T., and A. Lupas. 2004. CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics **20**:3702-3704.

Gordon, C., G. McGurk, M. Wallace, and N. D. Hastie. 1996. A conditional lethal mutant in the fission yeast 26 S protease subunit mts3+ is defective in metaphase to anaphase transition. J Biol Chem **271**:5704-5711.

Guindon, S., F. Delsuc, J. F. Dufayard, and O. Gascuel. 2009. Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol **537**:113-137.

Gunkel, F. A., and H. G. Gassen. 1989. Proteinase K from *Tritirachium album* Limber. Characterization of the chromosomal gene and expression of the cDNA in *Escherichia coli*.

Eur J Biochem **179**:185-194.

Gupta, R., Q. K. Beg, and P. Lorenz. 2002. Bacterial alkaline proteases: molecular approaches and industrial applications. Appl Microbiol Biotechnol **59**:15-32.

Hedstrom L. 2002. An overview of serine proteases. Curr Protoc Protein Sci **21**:21.10.

Hu, G., and R. J. St. Leger. 2004. A phylogenomic approach to reconstructing the diversification of serine proteases in fungi. J Evol Biol **17**:1204-1214.

Hunter, S., R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats. 2009. InterPro: the integrative protein signature database. Nucleic Acids Res **37**:D211-215.

Jones, A. L., B. B. Quimby, J. K. Hood, P. Ferrigno, P. H. Keshava, P. A. Silver, and A. H. Corbett. 2000. SAC3 may link nuclear protein export to cell cycle progression. Proc Natl Acad Sci U S A **97**:3224-3229.

Jousson, O., B. Lechenne, O. Bontems, B. Mignon, U. Reichard, J. Barblan, M. Quadroni, and M. Monod. 2004. Secreted subtilisin gene family in *Trichophyton rubrum*. Gene **339**:79-88.

Katoh, K., K. Kuma, H. Toh, and T. Miyata. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res **33**:511-518.

Kominami, K., and A. Toh-e. 1994. Characterization of the function of the NIN1 gene product of *Saccharomyces cerevisiae*. Exp Cell Res **211**:203-211.

Kuwahara, K., M. Yoshida, E. Kondo, A. Sakata, Y. Watanabe, E. Abe, Y. Kouno, S. Tomiyasu, S. Fujimura, T. Tokuhisa, H. Kimura, T. Ezaki, and N. Sakaguchi. 2000. A novel nuclear phosphoprotein, GANP, is up-regulated in centrocytes of the germinal center and associated with MCM3, a protein essential for DNA replication. Blood **95**:2321-2328.

Le, S. Q., and O. Gascuel. 2008. An improved general amino acid replacement matrix. Mol Biol Evol **25**:1307-1320.

Letunic, I., and P. Bork. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics **23**:127-128.

Letunic, I., T. Doerks, and P. Bork. 2009. SMART 6: recent updates and new developments. Nucleic Acids Res **37**:D229-232.

Luo, X., and K. Hofmann. 2001. The protease-associated domain: a homology domain associated with multiple classes of proteases. Trends Biochem Sci **26**:147-148.

Mahon, P., and A. Bateman. 2000. The PA domain: a protease-associated domain. Protein Sci **9**:1930-1934.

Marchler-Bauer, A., and S. H. Bryant. 2004. CD-Search: protein domain annotations on the fly. Nucleic Acids Res **32**:W327-331.

McGowan, C. H. 2003. Regulation of the eukaryotic cell cycle. Prog Cell Cycle Res **5**:1-4.

Monod, M. 2008. Secreted proteases from dermatophytes. Mycopathologia **166**:285-294.

Nakayama, K. 1997. Furin: a mammalian subtilisin/Kex2p-like endoprotease involved in processing of a wide variety of precursor proteins. Biochem J **327 (Pt 3)**:625-635.

North, B. J., and E. Verdin. 2004. Sirtuins: Sir2-related NAD-dependent protein deacetylases. Genome Biol **5**:224.

Rawlings, N. D., F. R. Morton, C. Y. Kok, J. Kong, and A. J. Barrett. 2008. MEROPS: the peptidase database. Nucleic Acids Res **36**:D320-325.

Reddy, P. V., C. K. Lam, and F. C. Belanger. 1996. Mutualistic fungal endophytes express a proteinase that is homologous to proteases suspected to be important in fungal pathogenicity. Plant Physiol **111**:1209-1218.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**:1572-1574.

Saeki, K., M. Okuda, Y. Hatada, T. Kobayashi, S. Ito, H. Takami, and K. Horikoshi. 2000. Novel oxidatively stable subtilisin-like serine proteases from alkaliphilic *Bacillus* spp.: enzymatic properties, sequences, and evolutionary relationships. Biochem Biophys Res Commun **279**:313-319.

Seeger, M., C. Gordon, K. Ferrell, and W. Dubiel. 1996. Characteristics of 26 S proteases from fission yeast mutants, which arrest in mitosis. J Mol Biol **263**:423-431.

Sharpton, T. J., J. E. Stajich, S. D. Rounsley, M. J. Gardner, J. R. Wortman, V. S. Jordar, R. Maiti, C. D. Kodira, D. E. Neafsey, Q. Zeng, C. Y. Hung, C. McMahan, A. Muszewska, M. Grynberg, M. A. Mandel, E. M. Kellner, B. M. Barker, J. N. Galgiani, M. J. Orbach, T. N. Kirkland, G. T. Cole, M. R. Henn, B. W. Birren, and J. W. Taylor. 2009. Comparative genomic analyses of the human fungal pathogens *Coccidioides* and their relatives. Genome Res **19**:1722-1731.

Siezen, R. J., B. Renckens, and J. Boekhorst. 2007. Evolution of prokaryotic subtilases: genome-wide analysis reveals novel subfamilies with different catalytic residues. Proteins **67**:681-694.

Soding, J., A. Biegert, and A. N. Lupas. 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res **33**:W244-248.

Spatafora J. W., G. H. Sung, D. Johnson, C. Hesse, B. O'Rourke, M. Serdani, R. Spotts, F. Lutzoni, V. Hofstetter, J. Miadlikowska, V. Reeb, C. Gueidan, E. Fraker, T. Lumbsch, R. Lücking, I. Schmitt, K. Hosaka, A. Aptroot, C. Roux, A. N. Miller, D. M. Geiser, J. Hafellner, G. Hestmark, A. E. Arnold, B. Büdel, A. Rauhut, D. Hewitt, W. A. Untereiner, M. S. Cole, C. Scheidegger, M. Schultz, H. Sipman, C. L. Schoch. 2006. A five-gene phylogeny of

Pezizomycotina. Mycologia **96**:1018-28.

Sreedhar, L., D. Y. Kobayashi, T. E. Bunting, B. I. Hillman, and F. C. Belanger. 1999. Fungal proteinase expression in the interaction of the plant pathogen *Magnaporthe poae* with its host. Gene **235**:121-129.

Takei, Y., and G. Tsujimoto. 1998. Identification of a novel MCM3-associated protein that facilitates MCM3 nuclear localization. J Biol Chem **273**:22177-22180.

Wang, R. B., J. K. Yang, C. Lin, Y. Zhang, and K. Q. Zhang. 2006. Purification and characterization of an extracellular serine protease from the nematode-trapping fungus *Dactylella shizishanna*. Lett Appl Microbiol **42**:589-594.

Wheeler, D. L., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, J. Ostell, K. D. Pruitt, G. D. Schuler, M. Shumway, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. 2008. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res **36**:D13-21.

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol **18**:691-699.

Wlodawer, A., M. Li, A. Gustchina, H. Oyama, B. M. Dunn, and K. Oda. 2003. Structural and enzymatic properties of the sedolisin family of serine-carboxyl peptidases. Acta Biochim Pol **50**:81-102.

Yakoby, N., D. Beno-Moualem, N. T. Keen, A. Dinoor, O. Pines, and D. Prusky. 2001. *Colletotrichum gloeosporioides* pelB is an important virulence factor in avocado fruit-fungus interaction. Mol Plant Microbe Interact **14**:988-995.

Yan, L., and Y. Qian. 2009. Cloning and heterologous expression of SS10, a subtilisin-like protease displaying antifungal activity from *Trichoderma harzianum*. FEMS Microbiol Lett **290**:54-61.

Yang, J., X. Huang, B. Tian, M. Wang, Q. Niu, and K. Zhang. 2005. Isolation and characterization of a serine protease from the nematophagous fungus, *Lecanicillium psalliotae*, displaying nematicidal activity. Biotechnol Lett **27**:1123-1128.

Zdobnov, E. M., and R. Apweiler. 2001. InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics **17**:847-848.

Figure Legends

Figure 1 MEROPS and Siezen et. al. (Siezen, Renckens, and Boekhorst 2007) subtilase classification. The schema shows the relationships between all categories (old and novel) applied in the publication. Arrows depict the hierarchical relationships, objects not separated by arrows correspond to one level of classification. The image was prepared with Dia (http://projects.gnome.org/dia/).

Figure 2 2D CLANS clustering of 1100 S8 and S53 proteases obtained from iterative BLAST searches against the fungal subset of the NCBI NR and Pfam databases. New groups of S8A subtilisin-like serine proteases are identified by “N” (like “New”) before the group number. Table 1 summarizes characteristics of the groups and Supplementary Fig. 4 shows their taxonomic distribution. Osf - oxidatively stable alkaline serine protease.

Figure 3 Domain architecture of S8 and S53 serine proteases. Protease K architectures are represented by figures A and B. Pyrolisin and osf protease architectures are shown on C,D and E. Kexin architectures are represented by E,F and G. New group 2 architectures have different carbohydrate hydrolyzing domains at their N prime ends, as in like on schema I. *Magnaporthe grisea* sequence GI:145608536 described in the domain result section is depicted on schema J. The novel *Aspergillus terreus* (GI: 115388617) domain fusion with cyclin is presented in schema K. L: *Giberella zeae* protein GI:46117066. M: *Chaetomium globosum* GI:116182816. N: shows a typical S53 architecture whereas O. (*Giberella zeae* protein GI:46111169) and P. (*Aspergillus terreus* protein GI:115384808) present some unusual domain co-occurrences.

Figure 4 Phylogenetic tree of the Onygenales family proteinase K proteases. Maximum Likelihood analysis of a set of 103 proteases was carried out using the LG+G model. Approximate likelihood ratio test SH-like branch supports above 50% are shown. Species abbreviations: Asp. – *Aspergillus*, Art. – *Arthoderma*, Coc. – *Coccidioides*, His. - *Histoplasma,* Tri. - *Trichophyton*, Unc. – *Uncinocarpus*.

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. Conserved motifs were predicted with MEME (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.

Figure 1

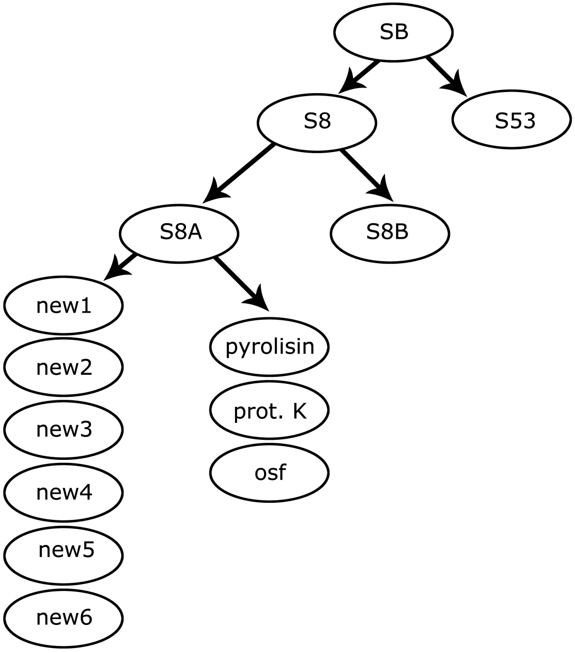

SB
S8
S53
S8A
S8B
new1
new2
new3
new4
new5
new6
pyrolisin
prot. K
osf

Figure 2

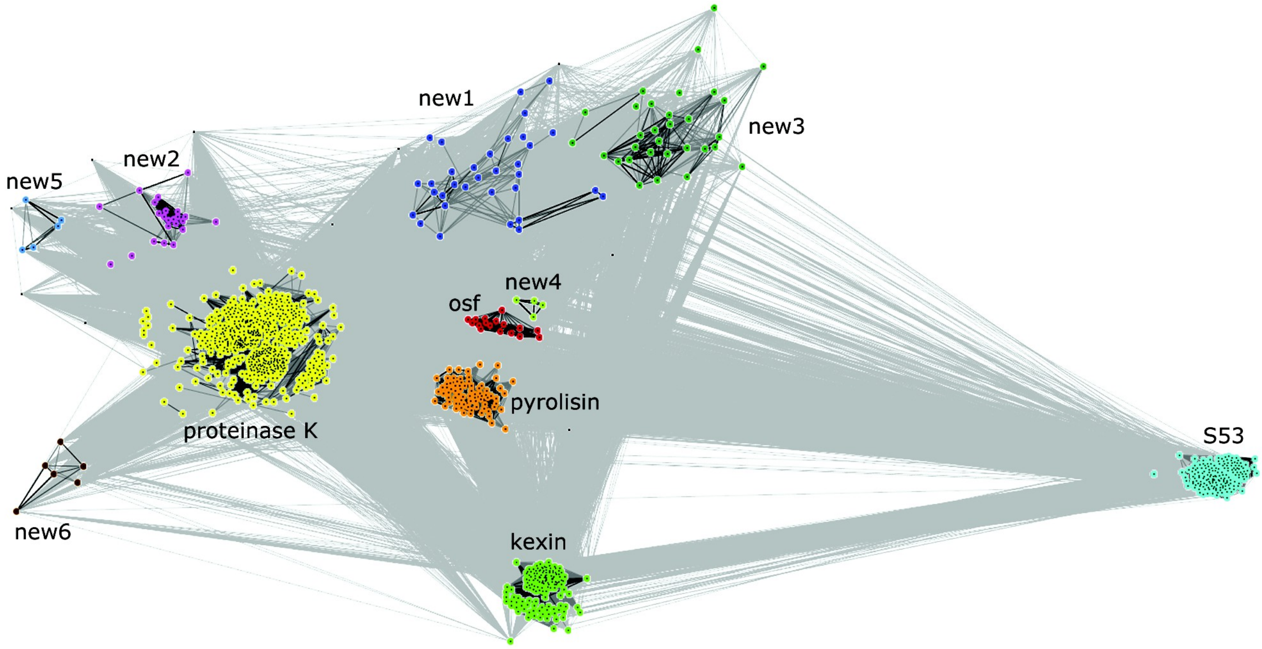
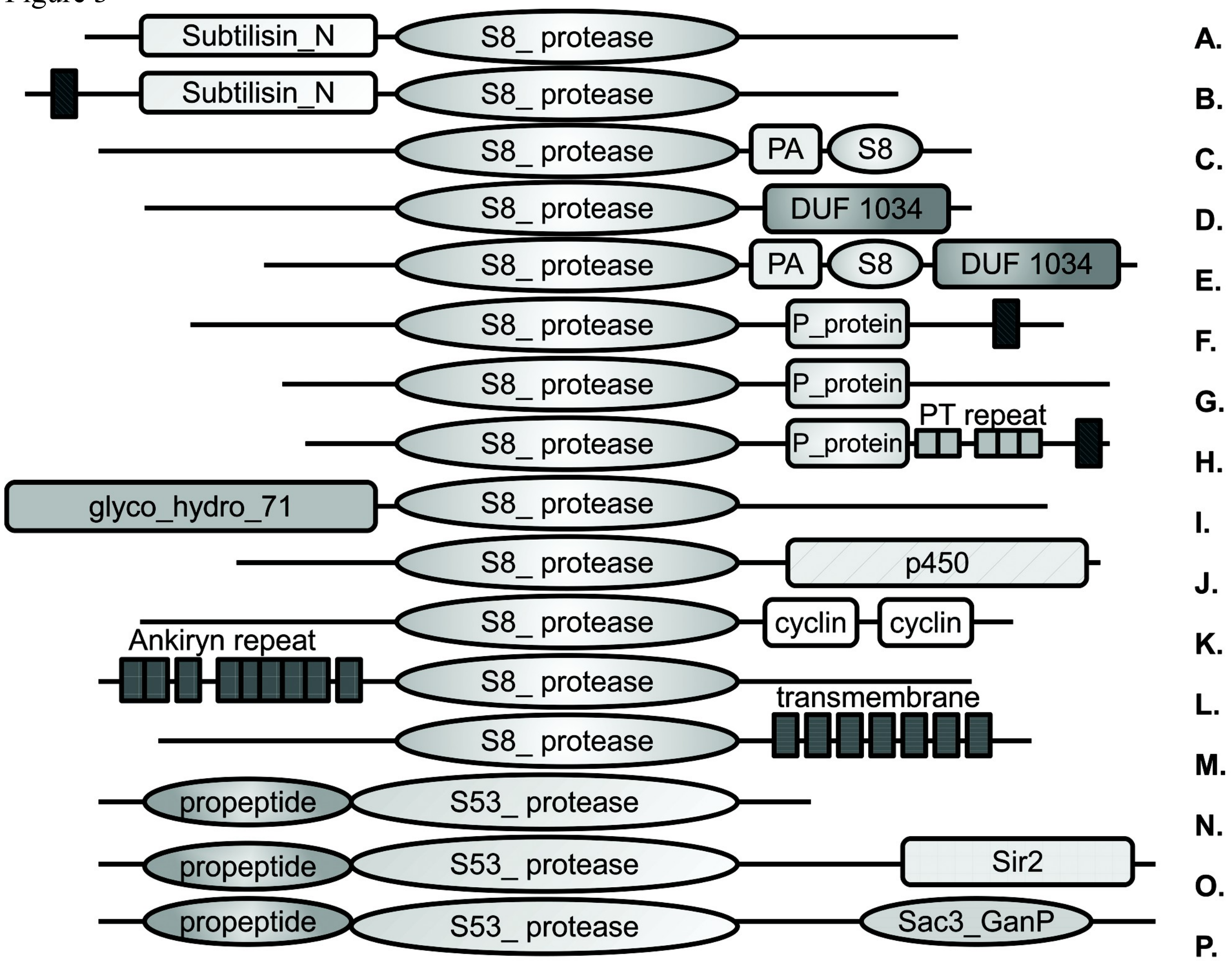
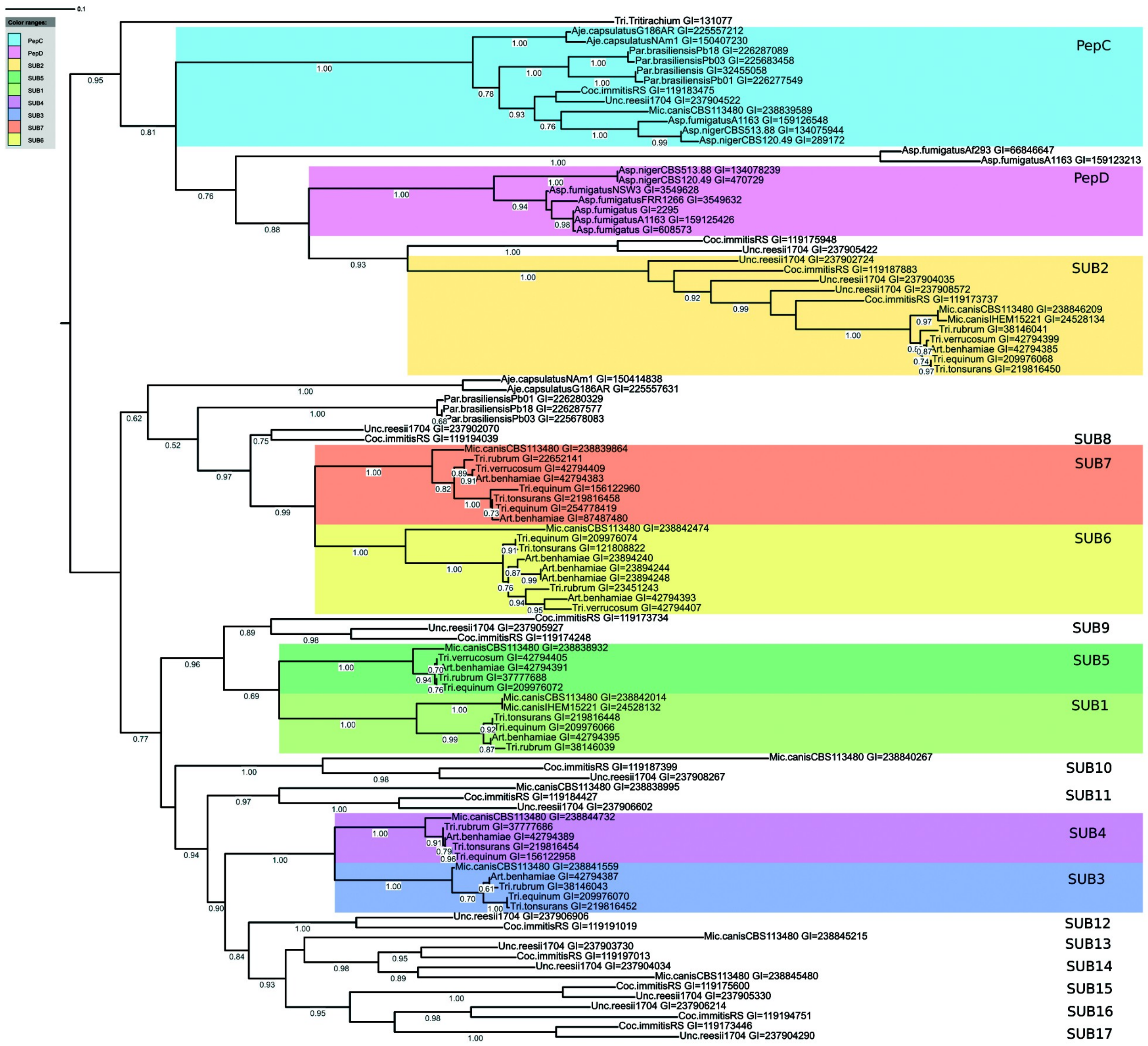

new1
new3
new2
new5
new4
osf
proteinase K
pyrolisin
S53
new6
kexin

Figure 3



Subtilisin_N
S8_ protease
A.
Subtilisin_N
S8_ protease
B.
S8_ protease
PA
S8
C.
S8_ protease
DUF 1034
D.
S8_ protease
PA
S8
DUF 1034
E.
S8_ protease
P_protein
F.
S8_ protease
P_protein
G.
PT repeat
S8_ protease
P_protein
H.
glyco_hydro_71
S8_ protease
I.
S8_ protease
p450
J.
S8_ protease
cyclin
cyclin
K.
Ankiryn repeat
S8_ protease
L.
transmembrane
S8_ protease
M.
propeptide
S53_ protease
N.
propeptide
S53_ protease
Sir2
O.
propeptide
S53_ protease
Sac3_GanP
P.

Figure 4


0.1
Color ranges:
PepC
PepD
SUB2
SUB5
SUB1
SUB4
SUB3
SUB7
SUB6
Tri.Tritirachium GI=131077
Aje.capsulatusG186AR GI=225557212
Aje.capsulatusNAm1 GI=150407230
Par.brasiliensisPb18 GI=226287089
Par.brasiliensisPb03 GI=225683458
Par.brasiliensis GI=32455058
Par.brasiliensisPb01 GI=226277549
Coc.immitisRS GI=119183475
Unc.reesii1704 GI=237904522
Mic.canisCBS113480 GI=238839589
Asp.fumigatusA1163 GI=159126548
Asp.nigerCBS513.88 GI=134075944
Asp.nigerCBS120.49 GI=289172
PepC
0.95
0.81
1.00
1.00
1.00
1.00
1.00
0.78
0.93
0.76
1.00
0.99
Asp.fumigatusAf293 GI=66846647
Asp.fumigatusA1163 GI=159123213
1.00
Asp.nigerCBS513.88 GI=134078239
Asp.nigerCBS120.49 GI=470729
Asp.fumigatusNSW3 GI=3549628
Asp.fumigatusFRR1266 GI=3549632
Asp.fumigatus GI=2295
Asp.fumigatusA1163 GI=159125426
Asp.fumigatus GI=608573
PepD
0.76
0.88
1.00
1.00
0.94
0.98
Coc.immitisRS GI=119175948
Unc.reesii1704 GI=237905422
Unc.reesii1704 GI=237902724
Coc.immitisRS GI=119187883
Unc.reesii1704 GI=237904035
Unc.reesii1704 GI=237908572
Coc.immitisRS GI=119173737
Mic.canisCBS113480 GI=238846209
Mic.canisIHEM15221 GI=24528134
Tri.rubrum GI=38146041
Tri.verrucosum GI=42794399
Art.benhamiae GI=42794385
Tri.equinum GI=209976068
Tri.tonsurans GI=219816450
SUB2
0.93
1.00
1.00
0.92
0.99
1.00
0.97
0.6
0.87
0.74
0.97
Aje.capsulatusNAm1 GI=150414838
Aje.capsulatusG186AR GI=225557631
Par.brasiliensisPb01 GI=226280329
Par.brasiliensisPb18 GI=226287577
Par.brasiliensisPb03 GI=225678083
Unc.reesii1704 GI=237902070
Coc.immitisRS GI=119194039
SUB8
1.00
1.00
0.68
0.62
0.52
0.75
0.97
0.99
Mic.canisCBS113480 GI=238839864
Tri.rubrum GI=22652141
Tri.verrucosum GI=42794409
Art.benhamiae GI=42794383
Tri.equinum GI=156122960
Tri.tonsurans GI=219816458
Tri.equinum GI=254778419
Art.benhamiae GI=87487480
SUB7
1.00
0.89
0.91
0.82
1.00
0.73
Mic.canisCBS113480 GI=238842474
Tri.equinum GI=209976074
Tri.tonsurans GI=121808822
Art.benhamiae GI=23894240
Art.benhamiae GI=23894244
Art.benhamiae GI=23894248
Tri.rubrum GI=23451243
Art.benhamiae GI=42794393
Tri.verrucosum GI=42794407
SUB6
1.00
1.00
0.91
0.87
0.99
0.76
0.94
0.95
Coc.immitisRS GI=119173734
Unc.reesii1704 GI=237905927
Coc.immitisRS GI=119174248
SUB9
0.89
0.98
0.96
Mic.canisCBS113480 GI=238838932
Tri.verrucosum GI=42794405
Art.benhamiae GI=42794391
Tri.rubrum GI=37777688
Tri.equinum GI=209976072
SUB5
1.00
0.70
0.94
0.76
0.69
Mic.canisCBS113480 GI=238842014
Mic.canisIHEM15221 GI=24528132
Tri.tonsurans GI=219816448
Tri.equinum GI=209976066
Art.benhamiae GI=42794395
Tri.rubrum GI=38146039
SUB1
1.00
1.00
0.99
0.92
0.87
0.77
Mic.canisCBS113480 GI=238840267
Coc.immitisRS GI=119187399
Unc.reesii1704 GI=237908267
SUB10
1.00
0.98
Mic.canisCBS113480 GI=238838995
Coc.immitisRS GI=119184427
Unc.reesii1704 GI=237906602
SUB11
0.97
1.00
Mic.canisCBS113480 GI=238844732
Tri.rubrum GI=37777686
Art.benhamiae GI=42794389
Tri.tonsurans GI=219816454
Tri.equinum GI=156122958
SUB4
1.00
0.91
0.79
0.96
0.94
1.00
Mic.canisCBS113480 GI=238841559
Art.benhamiae GI=42794387
Tri.rubrum GI=38146043
Tri.equinum GI=209976070
Tri.tonsurans GI=219816452
SUB3
1.00
0.61
0.70
1.00
0.90
Unc.reesii1704 GI=237906906
Coc.immitisRS GI=119191019
SUB12
1.00
Mic.canisCBS113480 GI=238845215
SUB13
Unc.reesii1704 GI=237903730
Coc.immitisRS GI=119197013
Unc.reesii1704 GI=237904034
SUB14
0.84
0.95
0.98
0.89
Mic.canisCBS113480 GI=238845480
Coc.immitisRS GI=119175600
Unc.reesii1704 GI=237905330
SUB15
0.93
1.00
Unc.reesii1704 GI=237906214
Coc.immitisRS GI=119194751
SUB16
0.95
0.98
Coc.immitisRS GI=119173446
Unc.reesii1704 GI=237904290
SUB17
1.00

CLANS clade	**D**TG	G**H**GTH	SGT**S**	**Additional domains**	**Taxonomic distribution**	**Number of sequences**
S8: new1	EP[VI][KR][IV]A[IV][LI]D[TS]G[IV]DxxHPY[IF]	HGT[HF][VI]AGL[LIV]LK[VL]AP[ND][AV][DE]	SGTS[VF][AS]TPIA[AV][GA][IL][AV]AN	WD40, ANK, TM, p-loop containing (PF00004 or PF05729)	*Pezizomycotina*	36
S8: new2	[EK]D[LF][GK]V[DG][EQ]FLIATEH[GD]CKNDGTGDNT[AG][DA]IN[SA]FLEKA	[DYW][PV]GP[AI][QR]P[DN][VL][EKR]HGT[GR] VASK[VI][LI]G[RA]NLG[SI]CQ	[SATLV][SADH]GTS[LY][AS]{PA][FVL][VLI][SA][GS][LV]	glyco_hydro_71 (PF03659) glyco_hydro_18 (PF00704), pectin lyase-like superfamily	*Pezizomycotina*	23
S8: new3	[PR][VI]KVA[LI]IDDG[VI]D	P[YW][YW]VSAxGHGTIMA[NR][ML]I[CL]R[IV][CN]PM	K[PS]VxYH[TS]GSS[VI][AS]TALAAGLA[AS]L[IV]LYCVR	ANK, cyclin(IPR006670)	*Pezizomycotina*	39
S8: new4	[KR][FY]P[DE][FY]DGR[GN]V[RTV]V[AG][IV]LDTGVDP[AG]A[AILP]GL	[LT]S[IL]V[TA][VL][SAC]G[TS]HGTHVAGI[IV]GA[HNQR][HT][PDQ][ED][HPQT]	LQ[NS][ST]QLMNGTSMSSP[NS]A[CA]G	low complexity	*Pezizomycotina*	4
S8: new5	GINA[RL]YAW[GT][FI][PT]GGDG[AL][GNR][TV][NGT]I[IV]D	[YFNW][YFPV][ADNRS]HGT[AS]V[LT]G[EAIQ][ML][LFG][MGQ][VAD][DV]N	[DW]Y[TY]DGF[SD]GTSGA[SA]PI[IV][VA]GAA[AL][AS]VQG	-	*Pezizomycotina* and *Taphrinomycotina*	8
S8: new6	DIP[AVI][YF]IVDTGAQ[IL]D[HN][PQ]	[NI]PHGT[GTA]	[VQS][QVE]GTS[VLE][AV][TW]	-	*Onygenales*	5
S8: osf	TEY[QT]GEGQV[VI][AC][VA][ACG]DTGFD[IK]G[KSD]T[DT]D	DPDGHGTHV[CA]GS[VI]LG[DN]GES[KN][ST]M	DPQ[WY][MF][FY]L[AS]GTSMATPLVAGC[AVC]AV[VL]RE[SA]LVKNG[TV][EK]NP	DUF1043(PF06280), PA(PF02225), Inhibitor_I9 (PF05922)	*Pezizomycotina*	17
S8: pyrolisin	VDKL[HR]A[EQK]GI[TL]GKG[VI][KR][VI][AG][VI][IV]D[TS]G[IV]DYTHPALG	[DM]DCxGHGTHV[AS]GI[IV][AG][AG]	YAVLSGTSMA[TC]P[YL]VAG[VI]AAL[YL]I	PA(PF02225), DUF1034(PF06280)	*Basidiomycota, Pezizomycotina* and *Pichia*	75
S8: proteinase K	[IVL]D[TS]G[IV][RN][IV]THP[ED]F[EG]GRA	DGNGHGTH[VC]AG[TI]I[GA][GS]KT[YF]GVAK[KN][AV]N[LI][VI]AVKV	SGTSMA[AST]PHVAGLAAYL[LM][SA]LEG	Inhibitor_I9 (PF05922), Cytochrome P450(PF00067)	all *Fungi*	621
S8: kexin	VDDGLD[YM][ET][SN]EDL[KA][DP]N[FY][NF]AEGS	[YW]DFND[NH]Tx[LDE]PKPRLSDDYHGTRCAGE[IV]AA	TNC[TS][TS]QH[GS]GTSA[AS][TA][PA][IL]AAG[IV]IAL[VA]	P domain (PF01483), TM	all *Fungi*	159
	Exxx**D**	SG**D**S	SGT**S**			
S53	E[AGS][NSTD]]LD[LV][QE]Y[AI]VG[LIV]SYP[LQ]PVT[EYL][YF][SQT]	V[IL]S[TI]SYG[ED][DN]EQS[VL]PxSYAxR[VQ]CN[EL][FY][AG][QK]LG[AL][RQ]GV[ST][VI][LI]F[SA]SGDSG	GxxxLVGGTSA[SA][AST]P[VT]FA[AS][IV][IV]AL[LI]N[DE][AE]	Pro-kuma_activ (SM00944), Sir2 (PF02146), sac_ganp (PF03399)	*Basidiomycota* and *Pezizomycotina*	190

Table 1