Journal of Microbiology (2020) Vol. 58, No. 9, pp. 805–811
DOI 10.1007/s12275-020-0288-9
eISSN 1976-3794
pISSN 1225-8873

Evolutionary analysis and protein family classification of chitin deacetylases in *Cryptococcus neoformans*[§]

Seungsue Lee, Hyun Ah Kang[*], and Seong-il Eyun[*]

Department of Life Science, Chung-Ang University, Seoul 06974, Republic of Korea

(Received Jun 2, 2020 / Revised Jul 15, 2020 / Accepted Jul 23, 2020)

***Cryptococcus neoformans* is an opportunistic fungal pathogen causing cryptococcal meningoencephalitis. Interestingly, the cell wall of *C. neoformans* contains chitosan, which is critical for its virulence and persistence in the mammalian host. *C. neoformans* (H99) has three chitin deacetylases (CDAs), which convert chitin to chitosan. Herein, the classification of the chitin-related protein (CRP) family focused on cryptococcal CDAs was analyzed by phylogenetics, evolutionary pressure (d_N/d_S), and 3D modeling. A phylogenetic tree of 110 CRPs revealed that they can be divided into two clades, CRP I and II with bootstrap values (> 99%). CRP I clade comprises five groups (Groups 1–5) with a total of 20 genes, while CRP II clade comprises sixteen groups (Groups 6–21) with a total of 90 genes. CRP I comprises only fungal CDAs, including all three *C. neoformans* CDAs, whereas CRP II comprises diverse CDAs from fungi, bacteria, and amoeba, along with other carbohydrate esterase 4 family proteins. All CDAs have the signal peptide, except those from group 11. Notably, CDAs with the putative *O*-glycosylation site possess either the glycosylphosphatidylinositol (GPI)-anchor motif for CRP I or the chitin-binding domain (CBD) for CRP II, respectively. This evolutionary conservation strongly indicates that the *O*-glycosylation modification and the presence of either the GPI-anchor motif or the chitin-binding domain is important for fungal CDAs to function efficiently at the cell surface. This study reveals that *C. neoformans* CDAs carrying GPI anchors have evolved divergently from fungal and bacterial CDAs, providing new insights into evolution and classification of CRP family.**

Keywords: *Cryptococcus neoformans*, chitin deacetylase, protein family classification, glycosylphosphatidylinositol-anchor, chitin binding domain, *O*-glycosylation site

Introduction

Chitin is the most abundant biopolymer in nature after cellulose (Sharp, 2013). It is present in vertebrates, fungi, and bacteria, but not in higher plants (Sharp, 2013; Tang *et al.*, 2015; Patel and Goyal, 2017) and is an important structural component of fungal cell walls and the exoskeletons of insects and crustaceans (Peter, 2005; Hoell *et al.*, 2010). This polymer has a unit called N-acetyl-D-glucosamine (GlcNAc) bound through β-1,4-linkages (Kumar, 2000). Chitosan, the deacetylated version of chitin, is generated by the conversion of GlcNAc to D-glucosamine via the enzymatic action of chitin deacetylases (CDAs) and thus possesses differing degrees of deacetylation (Rinaudo, 2006; Cheung *et al.*, 2015; Kyzas and Bikiaris, 2015). Chitosan is present in the cell wall and is important for cell wall integrity and spore formation in plant pathogenic and biocontrol fungi (Coluccio and Neiman, 2004; Palma-Guerrero *et al.*, 2008).

Several enzymes are involved in the biosynthesis and processing of chitin and chitosan (Supplementary data Fig. S1). Chitinase (EC 3.2.1.14) and chitosanase (EC 3.2.1.132) perform the endo-hydrolysis of chitin and chitosan, respectively (Jaworska, 2012; Lombard *et al.*, 2014; Kaczmarek *et al.*, 2019). β-N-Acetylhexosaminidase (EC 3.2.1.52) and exo-β-D-glucosaminidase (EC 3.2.1.165) (GlcNase) are exo-hydrolases associated with chitin and chitosan, respectively (Lombard *et al.*, 2014; Thadathil and Velappan, 2014; Kaczmarek *et al.*, 2019). The chitin-active lytic polysaccharide monooxygenase (EC 1.14.99.53) (LPMO) is capable of cleaving glycolic bonds in crystalline chitin by oxidizing either C1 or C4 of the glucopyranose ring (Lombard *et al.*, 2014; Courtade and Aachmann, 2019; Kaczmarek *et al.*, 2019). Chitin synthase (EC 2.4.1.16) utilizes UDP-N-acetylglucosamine to form the chitin polysaccharide (Arakane *et al.*, 2012; Lombard *et al.*, 2014). CDA (EC 3.5.1.41) induces the conversion of chitin and belongs to the carbohydrate esterase 4 (CE4) family (Lombard *et al.*, 2014; Kaczmarek *et al.*, 2019). The CE4 family is composed mainly of CDAs (EC 3.5.1.41) and chitooligosaccharide deacetylases (EC 3.5.1.-), peptidoglycan N-acetylglucosamine deacetylases (EC 3.5.1.104), peptidoglycan N-acetylmuramic acid deacetylases (EC 3.5.1.-), and poly-β-1,6-N-acetylglucosamine deacetylase (EC 3.5.1.-), though it also contains some acetylxylan esterase (EC 3.1.1.72) (Aragunde *et al.*, 2018).

CDAs (EC 3.5.1.41) are found in fungi, bacteria, one protozoan species (*Entamoeba histolytica*) and a few insects (*Tribolium castaneum*, *Drosophila melanogaster*, *Anopheles gambiae*, and *Apis mellifera*) (Dixit *et al.*, 2008; Zhao *et al.*, 2010; Jaworska, 2012; Grifoll-Romero *et al.*, 2018). CDAs perform critical roles in plant-pathogen interactions (Hoell *et al.*, 2010;

[*]For correspondence. (H.A. Kang) Email: hyunkang@cau.ac.kr; Tel.: +82-2-820-5863 / (S. Eyun) E-mail: eyun@cau.ac.kr; Tel.: +82-2-820-5163

[§]Supplemental material for this article may be found at http://www.springerlink.com/content/120956.

Cord-Landwehr *et al.*, 2016). Plant pathogens invade the host by secreting CDAs to eliminate the substrate of the chitinase (Cord-Landwehr *et al.*, 2016). The bacterial CDAs, also known as chitin oligosaccharide deacetylases, are found in *Vibrio*, *Shewanella*, and *Arthrobacter* (Kadokura *et al.*, 2007; Hirano *et al.*, 2015; Tuveng *et al.*, 2017; Grifoll-Romero *et al.*, 2018). Fungal CDAs are important for cell wall formation and integrity, defense mechanisms, fungal nutrition, morphogenesis, development, spore formation, germline adhesion, and fungal autolysis (Grifoll-Romero *et al.*, 2018). For example, CDAs are involved in spore formation and cell wall integrity in *Saccharomyces cerevisiae* (Lin *et al.*, 2013). The fungal CDAs were classified based on whether they play a role in the cell wall integrity and infection-associated autolysis using phylogenetic analysis (Grifoll-Romero *et al.*, 2018). However, three CDAs in *Cryptococcus neoformans* var. *grubii* serotype A (strain H99) were involved in both cell wall function and infection (Baker *et al.*, 2007; Grifoll-Romero *et al.*, 2018; Upadhya *et al.*, 2018). Thus, the functional classification of the two types of CDAs is not clear.

Cryptococcus neoformans is an opportunistic fungal pathogen causing cryptococcal meningoencephalitis in immunocompromised individuals. Notably, the cell wall of *C. neoformans* contains a substantial amount of chitosan. The chitin content in the cell wall of yeast is typically 1–2% (Klis, 1994; Banks *et al.*, 2005), while chitosan is not present in *S. cerevisiae*, *Candida*, or *Aspergillus* (Cabib *et al.*, 2001; Garcia-Rubio *et al.*, 2020). In *C. neoformans*, the cell wall is composed of chitin, glucans, melanin, and chitosan (Baker *et al.*, 2007; Garcia-Rubio *et al.*, 2020), and chitin and chitosan contents are 5% (Reiss *et al.*, 1986; Simmons, 1989). *C. neoformans* (H99) has three genes coding for CDAs (Cda1, Cda2, and Cda3), which convert chitin to chitosan by the hydrolysis of the acetamido group of GlcNAc. A chitosan-deficient *C. neoformans* strain was generated by deleting all three CDAs, and this strain is avirulent in mice, as it was rapidly cleared from the lungs of infected mice (Upadhya *et al.*, 2016). Here, we carried out the evolutionary analysis and protein family classification of cryptococcal CDAs based on bioinformatic approaches. Our analyses suggested that the three *C. neoformans* CDAs diverged from most fungal and bacterial CDAs, with distinctive structural organization. This study is expected to serve as a critical starting point for fungal protein family classification of chitin-related genes and understanding their evolutionary history.

Materials and Methods

Multiple sequence alignments and phylogenetic analysis

Multiple alignments of all chitin-related protein (CRP) sequences were generated using MAFFT with the L-INS-i algorithm and MAFFT-profile alignment option (ver. 7.407) (Katoh and Standley, 2013). Sequence alignment, the secondary structure of MP98, and active sites are displayed in Supplementary data Fig. S2. All CRP sequences used in this study are available at: http://eyunlab.cau.ac.kr/fungi_CRP.

Phylogenetic relationships were reconstructed by the maximum-likelihood method with the substitution model (JTT matrix) using the HPC-PTHREADS-AVX version of RAxML (ver. 8.2.12) (Stamatakis, 2014). The neighbor-joining phylogenetic method was performed using the Phylip package (ver. 3.697) (Felsenstein, 2005). Non-parametric bootstrapping with 1,000 pseudo-replicates was used to estimate the confidence level of branching topology for the maximum-likelihood and neighbor-joining phylogenies. The presentation of the phylogenies was generated with FigTree (ver. 1.4.3) (http://tree.bio.ed.ac.uk/software/figtree).

Domain prediction and protein sequence analysis for the protein diagram construction

Protein domain prediction was performed by Pfam (https://pfam.xfam.org), NCBI CDS (Conserved Domain Search, https://www.ncbi.nlm.nih.gov/Structure/cdd), and SMART (Simple Modular Architecture Research Tool, http://smart.embl-heidelberg.de) (Finn *et al.*, 2014; Letunic and Bork, 2018; Lu *et al.*, 2020). The presence of signal peptide, GPI-anchor, and location of the ω site for the GPI-anchor were predicted by SignalP (ver. 4.1) (http://www.cbs.dtu.dk/services/SignalP-4.1) and PredGPI (http://gpcr.biocomp.unibo.it/predgpi) (Pierleoni *et al.*, 2008; Petersen *et al.*, 2011). Transmembrane regions were predicted using TMHMM (ver. 2.0) (http://www.cbs.dtu.dk/services/TMHMM) (Krogh *et al.*, 2001). Potential *O*-glycosylation sites were predicted using NetOGlyc (ver. 4.0) (http://www.cbs.dtu.dk/services/NetOGlyc) (Steentoft *et al.*, 2013). We defined this site as one showing more than 50% of putative *O*-glycosylation sites, as assessed by NetOGlyc (ver.4.0) (Steentoft *et al.*, 2013). The protein diagrams were summarized from the most overlapping proteins or from the smallest proteins among those with similar structures, except for proteins shorter than 100 aa.

The calculation of amino acid composition ratio and support vector machine (SVM) analysis

The amino acid composition was calculated by COPID (COmposition-based Protein IDentification, http://crdd.osdd.net/raghava/copid) (Kumar *et al.*, 2008). The amino acid composition in the main domain (catalytic domain, for example, polysaccharide deacetylase) in the protein sequences was calculated. To calculate the amino acid composition ratio between CRP clades I and II, protein sequences without the main domain were obtained by removing the domain regions. The chitin-binding domain sequences were obtained in the Pfam alignments and then, their amino acid compositions were calculated (Finn *et al.*, 2014).

We used 3 packages: protr, kernlab, and RColorBrewer in R (Karatzoglou *et al.*, 2004; Neuwirth, 2014; Xiao *et al.*, 2015). Four steps were performed using the RColorBrewer function in R package: 1) readFASTA function [protein sequences were imported], 2) protcheck function [checking the amino acid types for excluding amino acids that were not among the 20 default amino acid types], 3) extractAAC function [amino acid composition was calculated], and 4) SVM graph was constructed with the ksvm function in the kernlab package and graph color selection.

Tests for positive selection

To examine positive selection, our codon alignment files and phylogenetic tree files were submitted to the CODEML program in PAML (ver. 4.9) (http://abacus.gene.ucl.ac.uk/software/paml.html). CODEML was used to analyze the data under six different models and parameters (Yang, 2007; Xu and Yang, 2013). The results are summarized in Supplementary data Table S3.

3D protein model prediction

The homology-based modeling of the group 4 proteins was performed using the SWISS-MODEL Web server (Waterhouse *et al.*, 2018). The template was selected from the results obtained from SWISS-MODEL (Waterhouse *et al.*, 2018). Structural modeling of Cda2 was conducted using Phyre2 (intensive mode) (Kelley *et al.*, 2015). The graphical representation of the template and model protein structure was prepared with PyMOL (ver. 2.0) (DeLanoScientific). The secondary structure was predicted by PSIPRED (ver. 4.0) (Buchan and Jones, 2019). We confirmed the secondary structure via additional alignment with group 4 proteins included in Cda2 and their template, *Colletotrichum lindemuthianum* CDA (PDB ID: 2IW0).

Results and Discussion

Collection of the CRP family

To perform the classification of chitin-related protein (CRP) family focused on *C. neoformans* CDAs, we collected 110 genes involved in the synthesis and processing of chitin and chitosan from fungi, bacteria, and amoeba; these are designated as members of the CRP family. Three representative CDA sequences (Cda1 [XP_012050538.1], Cda2 [XP_012049402.1, MP98], and Cda3 [XP_012049409.1, MP84]) in *C. neoformans* var. *grubii* serotype A (strain H99) were obtained from four previous studies (Levitz *et al.*, 2001; Biondo *et al.*, 2005; Loftus *et al.*, 2005; Baker *et al.*, 2007). Furthermore, CDAs and other CE4 proteins from diverse species such as fungi, bacteria, and amoeba that were collected from six studies (Levitz *et al.*, 2001; Biondo *et al.*, 2005; Levitz and Specht, 2006; Baker *et al.*, 2007; Aragunde *et al.*, 2018; Grifoll-Romero *et al.*, 2018) were added to our sequence sets. A similarity search was carried out using stand-alone blast (tblastn and blastp) against multiple species with an E-value threshold of 3.63×10^{-71} (Supplementary data Table S1). Additional protein sequence similarity searches using web-blast and PDB information were also performed (Berman *et al.*, 2000; Camacho *et al.*, 2009). To exclude proteins unrelated to CRPs in the previous processes, we confirmed whether they belong to the CE4 gene family by analysis using the Pfam and CAZy databases (Finn *et al.*, 2014; Lombard *et al.*, 2014). Our dataset was clustered at 95% identity using by USEARCH (ver. 11.0.667) (Edgar, 2010). A total of 110 sequences were obtained; they are summarized in Table 1 (46 fungi [26 genera], 25 bacteria [19 genera], and 1 amoeba; the species used are summarized in Supplementary data Table S4) (Levitz *et al.*, 2001; Biondo *et al.*, 2005; Levitz and Specht, 2006; Baker *et al.*, 2007; Aragunde *et al.*, 2018; Grifoll-Romero *et al.*, 2018). All proteins were categorized into CDAs and the CE4 proteins.

Table 1. The number of protein sequences obtained from each procedure

Source	Sequence number
Aragunde *et al.* (2018)	15
Baker *et al.* (2007)	3
Biondo *et al.* (2005)	1
Grifoll-Romero *et al.* (2018)	14
Levitz *et al.* (2001)	1
Web-blast	55
Similarity search by stand-alone blast[a]	10
PDB information[b]	11
Total	110

[a] Stand-alone blast search against in the genomes or proteomes

[b] In the PDB site (www.rcsb.org), we confirmed the existence of additional protein sequences that shared the same PDB in other organisms.

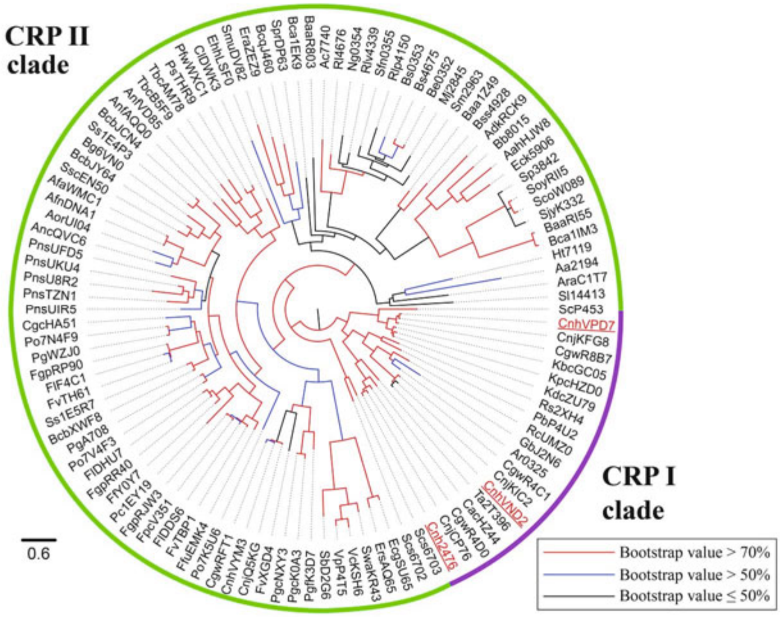

Fig. 1. The maximum-likelihood phylogeny of chitin deacetylase and other carbohydrate esterase 4 family proteins. The collected 110 CRP genes from fungi, bacteria, and amoeba were subjected to phylogenetic analysis. The purple and green arcs indicate the CRP I clade and CRP II clade, respectively. The CRP II clade is composed of chitin deacetylase and other carbohydrate esterase 4 family proteins. Non-parametric bootstrapping with 1,000 pseudo-replicates was used to estimate the confidence of branching topology for the maximum-likelihood and neighbor-joining phylogenies. The blue node indicates that the bootstrap value of the node is supported by 50% and the red node indicates that the bootstrap value of the node is supported by 70%. The black node indicates that the bootstrap value of the node is not supported because the value is lower than 50%. CDAs from *C. neoformans* (H99) are red-colored and underlined. The accession numbers are listed in Supplementary data Table S2.

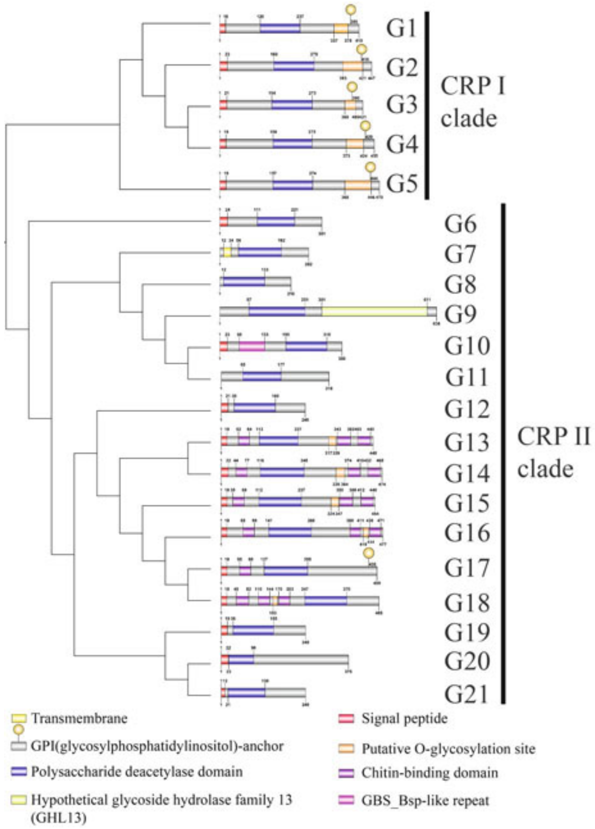
G1
G2
G3
G4
G5
CRP I clade
G6
G7
G8
G9
G10
G11
G12
G13
G14
G15
G16
G17
G18
G19
G20
G21
CRP II clade
Transmembrane
GPI(glycosylphosphatidylinositol)-anchor
Polysaccharide deacetylase domain
Hypothetical glycoside hydrolase family 13 (GHL13)
Signal peptide
Putative O-glycosylation site
Chitin-binding domain
GBS_Bsp-like repeat

Fig. 2. Phylogenetic analysis of the protein family group and domain diagram. The representative proteins in the diagram are the most overlapping structures or the smallest proteins among those showing similar structures. The protein diagrams summarize the domain organization of the representative protein. CRP I clade proteins have four domains (signal peptide, polysaccharide deacetylase, putative *O*-glycosylation site, and GPI anchor), and CRP II clade proteins contain common domain (polysaccharide deacetylase) and additional diverse structures (1 to 7 regions, signal peptide, transmembrane region, hypothetical glycoside family 13, GBS_Bsp-like repeat region, chitin-binding domain, putative *O*-glycosylation site and GPI anchor). The phylogenetic relationship is based on Fig. 1.

Phylogenetic analysis of CRPs

The phylogenetic tree is largely divided into two clades, CRP I and CRP II (Fig. 1, bootstrap values > 99%). The CRP I clade can be further divided into five groups (groups 1 to 5) and the CRP II clade has sixteen groups (groups 6 to 21) (Figs. 1 and 2). All groups are strongly supported by a high bootstrap value (> 61.7%). The CRP I clades have only fungal CDAs and are supported with a high bootstrap value (> 70%). The CRP II clades are composed of CE4 proteins and have various bootstrap values (< 50–100%) because their proteins are probably from more diverse species.

Protein diagram construction and the analysis of the structure of each CDA group by domain and site prediction

To identify the structural characteristics between two clades, several bioinformatic tools were used to analyze protein domains (Pfam, NCBI CDS search, and SMART) and specific sites (SignalP, PredGPI, TMHMM, and NetOGlyc). The common structures in each group are summarized in Fig. 2. Proteins from the CRP I clade (groups 1 to 5) showed an identical structural organization; they comprised a signal peptide, polysaccharide deacetylase (PDA) domain, the putative *O*-glycosylation site, and the GPI-anchor. None of the members of CRP I clade has the chitin-binding domain (CBD). In contrast, all the CDAs (groups 1 to 6, 11 to 19, 21) showed the structural features containing the signal peptide, PDA domain, the putative *O*-glycosylation site, and CBD except for CDAs from group 11, which have only the PDA domain. In the CRP II clade, CDAs were found in eleven groups (groups 6, 11 to 19, 21). In all the CDAs, the putative *O*-glycosylation site is generally located between the catalytic domain and the GPI-anchor or the CBD (Cord-Landwehr *et al.*, 2016; Hoßbach *et al.*, 2018). In general, the GPI-anchor is accompanied with a serine/threonine-rich region (de Groot *et al.*, 2003; González *et al.*, 2012) as observed in all the members from CRP I. However, CDAs from group 17 in CRP II are predicted to have only a GPI-anchor without the putative *O*-glycosylation site. It is interesting that all CBD-containing CDAs from CRP II also possessed the putative *O*-glycosylation site, indicating that the modification by *O*-glycosylation might be important for fungal CDAs to perform their function at the cell surface.

Group 8 possesses only PDA domain, but all other groups has additional regions. For example, groups 7, 9, 10, and 20 have additional regions such as transmembrane region in group 7, hypothetical glycoside hydrolase family 13 in group 9, signal peptide and GBS Bsp-like repeat in group 10, and signal peptide in group 20. While groups 13 to 18 have CBDs, five groups (6, 11, 12, 19, and 21) do not have CBDs.

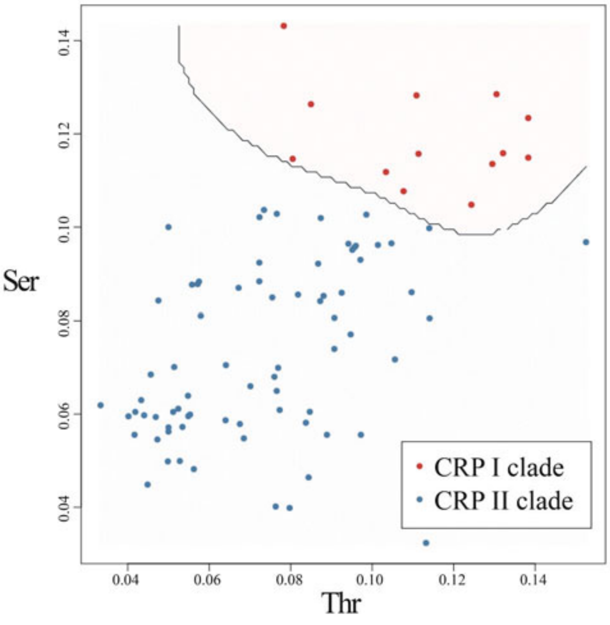

Fig. 3. Support vector machine (SVM) analysis of protein composition between CRP clade I and II. The amino acid composition was analyzed by protr package in R, and the SVM model is constructed using the kernlab package in R. The graph shows that clade I proteins showed a higher composition ratio of both serine and threonine than clade II proteins (Both serine and threonine, CRP I clade = 11%, CRP II clade = 7%).

Amino acid composition ratio calculation relative to the CBD

We confirmed whether there was a difference between the two clades with regard to the amino acid composition ratio. The amino acid composition ratios can be used as protein discrimination (Liu *et al.*, 2003; Nanni *et al.*, 2012) and also cryptococcal CDAs had generally Serine/Threonine-rich region. The CRP I clade proteins showed high serine and threonine composition (S/T) ratios of 10.76% and 10.55%. The CRP II clade proteins have S/T ratios of 7.62% and 7.44%. Among the CRP II clade, CDAs having two or more CBDs (26 proteins from nongroup and groups 13 to 18) showed S/T ratios similar to those of the CRP I clade, with ratios of 9.51% and 8.77%. To confirm the serine and threonine compositions as a suitable parameter for the presence of the CBD, SVM analysis was used to determine the difference in S/T ratios between the two clades (Fig. 3). The SVM is a machine learning algorithm created for the classification of data (Cortes and Vapnik, 1995; Burges, 1998). Therefore, the ratio of serine and threonine clarifies the difference between the two clades.

Motif analysis of CDAs

Previous studies identified five motifs and revealed their functions in fungal CDAs (Blair *et al.*, 2006; Aragunde *et al.*, 2018; Grifoll-Romero *et al.*, 2018). These five motifs were identified in *C. neoformans* Cda2. Among the motifs, those interacting with the metal cation and active site are indicated in Fig. 4 and Supplementary data Fig. S2. Among all CDAs in this study, the ratio of CDAs containing the conserved five motifs was 92.21%, while two groups (groups 11 and 21), representing *Encephalitozoon* and *Schizosaccharomyces* CDAs, have no motifs. The *Encephalitozoon* CDAs were reported as inactive enzymes with functions other than the deacetylase activity (Urch *et al.*, 2009; Aragunde *et al.*, 2018). The *Schizosaccharomyces* CDAs were determined to represent proteins more closely related to CE4 gene family than the CDA groups in our phylogenetic tree. Therefore, our phylogeny and motif analysis suggest the possibility that *Schizosaccharomyces* CDAs may have an additional function that is not deacetylation.

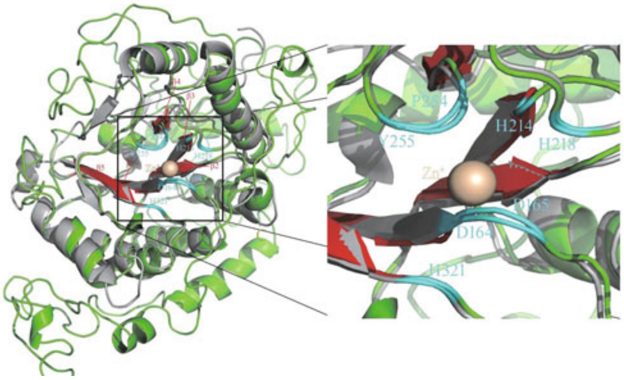

Fig. 4. The 3D structural model of Cda2 (green) superimposed with the fungal chitin deacetylase (CDA). The active sites and β-strands, which are indicated by the expansion of the active sites of in *C. neoformans* Cda2 are presented. The template for constructing the protein model above (*Colletotrichum lindemuthianum* CDA, PDB: 2IW0, AAT68493.1) was selected using SWISS-MODEL and is indicated using a gray color (http://swissmodel.expasy). The predicted 3D structure of Cda2 was modeled by Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2). The wheat sphere is Zn^{+} (zinc ion). The amino acids of the active site are indicated in cyan and labeled with one letter and the positions in the template and model protein. β-Strands in the 3D-structure are indicated in red and marked as β1–β5. The graphical representation was generated using PyMOL (ver. 2.0).

The 3D modeling and positive selection

An estimation of the d_N/d_S ratio, *i.e.*, the ratio between the nonsynonymous (d_N) and the synonymous (d_S) substitution rates in an alignment of amino acid-coding sequences (named "d_N/d_S methods" in the rest of the article) has been used extensively to identify individual codon positions evolving under positive selection (Nielsen, 2005; Moury and Simon, 2011). Synonymous substitution is thought to be largely neutral, while non-synonymous substitution is influenced by selection (Nielsen, 2005; Tennessen, 2008). Thus, when the d_N/d_S ratio is high between genes, functional divergence is assumed to be high, whereas when the d_N/d_S ratio is low, the functional properties of the gene products involved are thought to be conserved (Nielsen, 2005; Tennessen, 2008). To consider positive selection in CDAs and CRPs, the d_N/d_S ratio calculation was performed using PAML (ver. 4.9). We analyzed each group in our dataset and identified no groups, confirming that positive selection was partially present (Supplementary data Table S3).

When the d_N/d_S ratio is high between genes, functional divergence is assumed to be high (Nielsen, 2005; Tennessen, 2008). Group 4 has the higher d_N/d_S ratio (0.19951) among variable CRP I groups, this group was selected for the 3D modeling. To explore structure of group 4 proteins in more detail, we performed 3D modeling of *C. neoformans* Cda2. In the SWISS-MODEL, the template is *Colletotrichum lindemuthianum* CDA (PDB ID: 2IW0). Figure 4 shows the position of the β-strand and active sites. We paid attention to the β-strands because β-strands are composed of the section within the active site, which is well conserved between *C. neoformans* Cda2 and the template. The secondary structure in Cda2 and the active sites in group 4 and the template are shown in Supplementary data Fig. S1. The 3D modeling and alignment indicate that the foundational structure (active site and β-strands) is same between identified CDAs and representative cryptococcal CDA, Cda2.

Conclusion

Our study shows a clear structural difference between CRP clade I and II. In the CRP I clade, CDAs are composed of four domains (signal peptide, the catalytic domain, the putative *O*-glycosylation site, and GPI-anchor). In the CRP II clade, CDAs display two representative structural organizations: one with two domains (signal peptide and the catalytic domain) and the other with four domains (signal peptide, the catalytic domain, the putative *O*-glycosylation site, and CBD). It is noticeable that the putative *O*-glycosylation site is commonly conserved in all the CDAs not only in CRP clade I but also in CRP clade II, suggesting the importance of this post-translational modification for the function or structural stability of CDAs. Elucidating the specific roles of the putative

O-glycosylation site is further required. This study provides new insights into the protein family classification of chitin-related genes because three CDAs from *C. neoformans* H99 are clustered divergently from most fungal and bacterial CDAs because of their distinctive structural organization. Particularly, the possession of GPI-anchors instead of CBD, or the opposite, appears to be a major event causing the CRP I clade to diverge from the common ancestor of CRP.

Acknowledgments

This research was supported by the National Research Foundation of Korea (2018R1A5A1025077 and 2019R1A2C1084942) and the Chung-Ang University Research Scholarship Grants in 2018.

Conflict of Interest

The authors declare no conflicts of interest.

References

Aragunde, H., Biarnés, X., and Planas, A. 2018. Substrate recognition and specificity of chitin deacetylases and related family 4 carbohydrate esterases. *Int. J. Mol. Sci.* **19**, 1–30.

Arakane, Y., Taira, T., Ohnuma, T., and Fukamizo, T. 2012. Chitin-related enzymes in agro-biosciences. *Curr. Drug Targets* **13**, 442–470.

Baker, L.G., Specht, C.A., Donlin, M.J., and Lodge, J.K. 2007. Chitosan, the deacetylated form of chitin, is necessary for cell wall integrity in *Cryptococcus neoformans*. *Eukaryot. Cell* **6**, 855–867.

Banks, I.R., Specht, C.A., Donlin, M.J., Gerik, K.J., Levitz, S.M., and Lodge, J.K. 2005. A chitin synthase and its regulator protein are critical for chitosan production and growth of the fungal pathogen *Cryptococcus neoformans*. *Eukaryot. Cell* **4**, 1902–1912.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The protein data bank. *Nucleic Acids Res.* **28**, 235–242.

Biondo, C., Messina, L., Bombaci, M., Mancuso, G., Midiri, A., Beninati, C., Cusumano, V., Gerace, E., Papasergi, S., and Teti, G. 2005. Characterization of two novel cryptococcal mannoproteins recognized by immune sera. *Infect. Immun.* **73**, 7348–7355.

Blair, D.E., Hekmat, O., Schüttelkopf, A.W., Shrestha, B., Tokuyasu, K., Withers, S.G., and Van Aalten, D.M. 2006. Structure and mechanism of chitin deacetylase from the fungal pathogen *Colletotrichum lindemuthianum*. *Biochemistry* **45**, 9416–9426.

Buchan, D.W. and Jones, D.T. 2019. The PSIPRED protein analysis workbench: 20 years on. *Nucleic Acids Res.* **47**, W402–W407.

Burges, C.J.C. 1998. A tutorial on support vector machines for pattern recognition. *Data. Min. Knowl. Discov.* **2**, 121–167.

Cabib, E., Roh, D.H., Schmidt, M., Crotti, L.B., and Varma, A. 2001. The yeast cell wall and septum as paradigms of cell growth and morphogenesis. **276**, 19679–19682.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. 2009. BLAST+: architecture and applications. *BMC Bioinfomatics* **10**, 421.

Cheung, R.C.F., Ng, T.B., Wong, J.H., and Chan, W.Y. 2015. Chitosan: An update on potential biomedical and pharmaceutical applications. *Mar. Drugs* **13**, 5156–5186.

Coluccio, A. and Neiman, A.M. 2004. Interspore bridges: a new feature of the *Saccharomyces cerevisiae* spore wall. *Microbiology* **150**, 3189–3196.

Cord-Landwehr, S., Melcher, R.L., Kolkenbrock, S., and Moerschbacher, B.M. 2016. A chitin deacetylase from the endophytic fungus *Pestalotiopsis* sp. efficiently inactivates the elicitor activity of chitin oligomers in rice cells. *Sci. Rep.* **6**, 38018.

Cortes, C. and Vapnik, V. 1995. Support-vector networks. *Mach. Learn.* **20**, 273–297.

Courtade, G. and Aachmann, F.L. 2019. Chitin-active lytic polysaccharide monooxygenases. *In* Yang, Q. and Fukamizo, T. (eds.), Targeting Chitin-containing Organisms. Advances in Experimental Medicine and Biology, vol. 1142. Springer, Singapore.

de Groot, P.W.J., Hellingwerf, K.J., and Klis, F.M. 2003. Genome-wide identification of fungal GPI proteins. *Yeast* **20**, 781–796.

Dixit, R., Arakane, Y., Specht, C.A., Richard, C., Kramer, K.J., Beeman, R.W., and Muthukrishnan, S. 2008. Domain organization and phylogenetic analysis of proteins from the chitin deacetylase gene family of *Tribolium castaneum* and three other species of insects. *Insect Biochem. Mol. Biol.* **38**, 440–451.

Edgar, R.C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461.

Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6., Distributed by the author. Department of Genome Sciences, University of Washington, Seattle, USA.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., *et al.* 2014. Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230.

Garcia-Rubio, R., de Oliveira, H.C., Rivera, J., and Trevijano-Contador, N. 2020. The fungal cell wall: *Candida*, *Cryptococcus*, and *Aspergillus* species. *Front. Microbiol.* **10**, 2993.

González, M., Brito, N., and González, C. 2012. High abundance of Serine/Threonine-rich regions predicted to be hyper-*O*-glycosylated in the secretory proteins coded by eight fungal genomes. *BMC Microbiol.* **12**, 213.

Grifoll-Romero, L., Pascual, S., Aragunde, H., Biarnés, X., and Planas, A. 2018. Chitin deacetylases: Structures, specificities, and biotech applications. *Polymers* **10**, 352.

Hirano, T., Uehara, R., Shiraishi, H., Hakamata, W., and Nishio, T. 2015. Chitin oligosaccharide deacetylase from *Shewanella woodyi* ATCC51908. *J. Appl. Glycosci.* **62**, 153–157.

Hoßbach, J., Bußwinkel, F., Kranz, A., Wattjes, J., Cord-Landwehr, S., and Moerschbacher, B.M. 2018. A chitin deacetylase of *Podospora anserina* has two functional chitin binding domains and a unique mode of action. *Carbohydr. Polym.* **183**, 1–10.

Hoell, I.A., Vaaje-Kolstada, G., and Eijsink, V.G. 2010. Structure and function of enzymes acting on chitin and chitosan. *Biotechnol. Genet. Eng. Rev.* **27**, 331–366.

Jaworska, M.M. 2012. Kinetics of enzymatic deacetylation of chitosan. *Cellulose* **19**, 363–369.

Kaczmarek, M.B., Struszczyk-Swita, K., Li, X., Szczęsna-Antczak, M., and Daroch, M. 2019. Enzymatic modifications of chitin, chitosan, and chitooligosaccharides. *Front. Bioeng. Biotechnol.* **7**, 243.

Kadokura, K., Rokutani, A., Yamamoto, M., Ikegami, T., Sugita, H., Itoi, S., Hakamata, W., Oku, T., and Nishio, T. 2007. Purification and characterization of *Vibrio parahaemolyticus* extracellular chitinase and chitin oligosaccharide deacetylase involved in the production of heterodisaccharide from chitin. *Appl. Microbiol. Biotechnol.* **75**, 357–365.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. 2004. kernlab-An S4 package for kernel methods in R. *J. Stat. Softw.* **11**, 1–20.

Katoh, K. and Standley, D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780.

Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., and Sternberg, M.J.E. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858.

Klis, F.M. 1994. Review: cell wall assembly in yeast. *Yeast* **10**, 851–869.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580.

Kumar, M.R. 2000. A review of chitin and chitosan applications. *React. Funct. Polym.* **46**, 1–27.

Kumar, M., Thakur, V., and Raghava, G.P. 2008. COPid: composition based protein identification. *In Silico Biol.* **8**, 121–128.

Kyzas, G.Z. and Bikiaris, D.N. 2015. Recent modifications of chitosan for adsorption applications: a critical and systematic review. *Mar. Drugs* **13**, 312–337.

Letunic, I. and Bork, P. 2018. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* **46**, D493–D496.

Levitz, S.M., Nong, S., Mansour, M.K., Huang, C., and Specht, C.A. 2001. Molecular characterization of a mannoprotein with homology to chitin deacetylases that stimulates T cell responses to *Cryptococcus neoformans*. *Proc. Natl. Acad. Sci. USA* **98**, 10422–10427.

Levitz, S.M. and Specht, C.A. 2006. The molecular basis for the immunogenicity of *Cryptococcus neoformans* mannoproteins. *FEMS Yeast Res.* **6**, 513–524.

Lin, C.P., Kim, C., Smith, S.O., and Neiman, A.M. 2013. A highly redundant gene network controls assembly of the outer spore wall in *S. cerevisiae*. *PLoS Genet.* **9**, e1003700.

Liu, Q., Zhu, Y., Wang, B., and Li, Y. 2003. Identification of β-barrel membrane proteins based on amino acid composition properties and predicted secondary structure. *Comput. Biol. Chem.* **27**, 355–361.

Loftus, B.J., Fung, E., Roncaglia, P., Rowley, D., Amedeo, P., Bruno, D., Vamathevan, J., Miranda, M., Anderson, I.J., Fraser, J.A., *et al.* 2005. The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. *Science* **307**, 1321–1324.

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., and Henrissat, B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495.

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Marchler, G.H., Song, J.S., *et al.* 2020. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* **48**, D265–D268.

Moury, B. and Simon, V. 2011. dN/dS-based methods detect positive selection linked to trade-offs between different fitness traits in the coat protein of potato virus Y. *Mol. Biol. Evol.* **28**, 2707–2717.

Nanni, L., Brahnam, S., and Lumini, A. 2012. Wavelet images and Chou's pseudo amino acid composition for protein classification. *Amino Acids* **43**, 657–665.

Neuwirth, E. 2014. RColorBrewer: ColorBrewer palettes. *R package version* 1.1-2. *R J.*

Nielsen, R. 2005. Molecular signatures of natural selection. *Annu. Rev. Genet.* **39**, 197–218.

Palma-Guerrero, J., Jansson, H.B., Salinas, J., and Lopez-Llorca, L.V. 2008. Effect of chitosan on hyphal growth and spore germination of plant pathogenic and biocontrol fungi. *J. Appl. Microbiol.* **104**, 541–553.

Patel, S. and Goyal, A. 2017. Chitin and chitinase: role in pathogenicity, allergenicity and health. *Int. J. Biol. Macromol.* **97**, 331–338.

Peter, M.G. 2005. Chitin and chitosan in fungi. *In* Biopolymers Online: Biology·Chemistry·Biotechnology·Applications, Part 6,

Petersen, T.N., Brunak, S., von Heijne, G., and Nielsen, H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786.

Pierleoni, A., Martelli, P.L., and Casadio, R. 2008. PredGPI: a GPI-anchor predictor. *BMC Bioinfomatics* **9**, 392.

Reiss, E., White, E.H., Cherniak, R., and Dix, J.E. 1986. Ultrastructure of acapsular mutant *Cryptococcus neoformans* cap 67 and monosaccharide composition of cell extracts. *Mycopathologia* **93**, 45–54.

Rinaudo, M. 2006. Chitin and chitosan: properties and applications. *Prog. Polym. Sci.* **31**, 603–632.

Sharp, R.G. 2013. A review of the applications of chitin and its derivatives in agriculture to modify plant-microbial interactions and improve crop yields. *Agronomy* **3**, 757–793.

Simmons, R.B. 1989. Comparison of chitin localization in *Saccharomyces cerevisiae*, *Cryptococcus neoformans*, and *Malassezia* spp. *Mycol. Res.* **93**, 551–553.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313.

Steentoft, C., Vakhrushev, S.Y., Joshi, H.J., Kong, Y., Vester-Christensen, M.B., Schjoldager, K.T.B.G., Lavrsen, K., Dabelsteen, S., Pedersen, N.B., Marcos-Silva, L., *et al.* 2013. Precision mapping of the human *O*-GalNAc glycoproteome through SimpleCell technology. *EMBO J.* **32**, 1478–1488.

Tang, W.J., Fernandez, J., Sohn, J.J., and Amemiya, C.T. 2015. Chitin is endogenously produced in vertebrates. *Curr. Biol.* **25**, 897–900.

Tennessen, J.A. 2008. Positive selection drives a correlation between non-synonymous/synonymous divergence and functional divergence. *Bioinformatics* **24**, 1421–1425.

Thadathil, N. and Velappan, S.P. 2014. Recent developments in chitosanase research and its biotechnological applications: a review. *Food Chem.* **150**, 392–399.

Tuveng, T.R., Rothweiler, U., Udatha, G., Vaaje-Kolstad, G., Smalås, A., and Eijsink, V.G.H. 2017. Structure and function of a CE4 deacetylase isolated from a marine environment. *PLoS ONE* **12**, e0187544.

Upadhya, R., Baker, L.G., Lam, W.C., Specht, C.A., Donlin, M.J., and Lodge, K. 2018. *Cryptococcus neoformans* Cda1 and its chitin deacetylase activity are required for fungal pathogenesis. *mBio* **9**, e02087-18.

Upadhya, R., Lam, W.C., Maybruck, B., Specht, C.A., Levitz, S.M., and Lodge, J.K. 2016. Induction of protective immunity to cryptococcal infection in mice by a heat-killed, chitosan-deficient strain of *Cryptococcus neoformans*. *mBio* **7**, e00547-16.

Urch, J.E., Hurtado-Guerrero, R., Brosson, D., Liu, Z., Eijsink, V.G.H., Texier, C., and van Aalten, D.M.F. 2009. Structural and functional characterization of a putative polysaccharide deacetylase of the human parasite *Encephalitozoon cuniculi*. *Protein Sci.* **18**, 1197–1209.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., *et al.* 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303.

Xiao, N., Cao, D.S., Zhu, M.F., and Xu, Q.S. 2015. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* **31**, 1857–1859.

Xu, B. and Yang, Z. 2013. PAMLX: a graphical user interface for PAML. *Mol. Biol. Evol.* **30**, 2723–2724.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.

Zhao, Y., Park, R.D., and Muzzarelli, R.A. 2010. Chitin deacetylases: properties and applications. *Mar. Drugs* **8**, 24–46.