The Expansin Engineering Database: a navigation and classification tool for expansins and homologues

Caroline Lohoff[1], Patrick Buchholz[1], Marilize Le Roes - Hill[2], and Juergen Pleiss[1]

[1]University of Stuttgart
[2]Cape Peninsula University of Technology

April 28, 2020

Abstract

Expansins have the remarkable ability to loosen plant cell walls and cellulose material without showing catalytic activity and therefore have potential applications in biomass degradation. To support the study of sequence-structure-function relationships and the search for novel expansins, the Expansin Engineering Database (ExED, https://exed.biocatnet.de) collected sequence and structure data on expansins from Bacteria, Fungi, and Viridiplantae, and expansin-like homologues such as carbohydrate binding modules, glycoside hydrolases, loosenins, swollenins, cerato-platanins, and EXPNs. Based on global sequence alignment and protein sequence network analysis, the sequences are highly diverse. However, many similarities were found between the expansin domains. Newly created profile hidden Markov models of the two expansin domains enable standard numbering schemes, comprehensive conservation analyses, and genome annotation. Conserved key amino acids in the expansin domains were identified, a refined classification of expansins and carbohydrate binding modules was proposed, and new sequence motifs facilitate the search of novel candidate genes and the engineering of expansins.

Keywords

protein domains, sequence motifs, sequence analysis, standard positions, CBM63, GH45

Introduction

Expansins are plant cell wall loosening proteins without apparent catalytic activity, which have been identified in a broad range of organisms[1–4]. The loosening mechanism is still elusive, but it has been suggested that the non-covalent interactions between cellulose microfibrils are weakened and moved against each other, thus the tight cellulosic structure is loosened[1]. The interactions between expansins and the plant cell wall, which consists of lignin, hemicellulose, and cellulose, require further investigation[5]. Expansins were first discovered in plants and were described as proteins mediating pH-dependent extension and stress relaxation of cell walls[6]. Based on phylogenetic analysis, it has been proposed that expansins in *Bacteria* and Fungi resulted from multiple horizontal gene transfers from plants to microbes[7], but there is also the possibility that the microbial expansin subfamily evolved first in ancient marine microorganisms, and then diversified into distinct terrestrial plant subfamilies[8].

Expansins consist of two tightly packed protein domains, connected by a short linker and preceded by a signal peptide[9]**(Figure 1)** . Both expansin domains need to be connected for effective wall extension activity and weakening filter paper[10,11]. The C-terminal domain of EXLX1 (expansin-like X) from *Bacillus subtilis* dominates the binding to cellulose and to matrix polysaccharides of cell walls through electrostatic or polar interaction[10]. The *Zea mays* β-expansin (*Zm* EXPB1) primarily binds glucuronoarabinoxylan, the major matrix polysaccharide in grass cell walls, and loosens it[12].

Key amino acids in the N-terminal domain of *Bacillus subtilis*expansin-like protein 1 (*Bs* EXLX1) are two threonines at positions 12 and 14, a serine at position 16, two aspartates at positions 71 and 82, a tyrosine at position 73, and a glutamic acid at position 75[10], numbered according to[13]. The threonine at standard position 12 is strongly conserved, but not essential for activity[10]. The aspartate at position 82 is crucial for activity; the threonine at position 14, the aspartate at position 71, and the tyrosine at position 73 are important for activity; and the serine at position 16 and the glutamic acid at position 75 play moderate roles in wall creep activity[10]. Three disulfide bridges can be found in the N-terminal domain of*Zm* EXPB1[14], and the six participating cysteines are highly conserved in the plant expansin groups, EXPA (expansin A) and EXPB (expansin B)[14]. An additional highly conserved cysteine pair is considered as a fourth disulfide bridge in plant α-expansins[15]. In the expansin protein *Sc* Exlx1 from the Basidiomycete fungus *Schizophylum commune* , three disulfide bonds are predicted[16], whereas there is a lack of disulfide bridges in *Bs* EXLX1[13] and many other bacterial expansins.

The N-terminal expansin domain is formed by a six-stranded double- β-barrel[13] that is shared by several protein superfamilies[17], e.g. glycoside hydrolase family 45 (GH45)[18,19]. The expansin-like proteins found in Fungi such as loosenins, EXPNs, or cerato-platanins are single-domain proteins that resemble the N-terminal domain of expansins[20–22].

The C-terminal expansin domain is responsible for the binding to cellulosic material and is formed by two stacked β-sheets with an immunoglobulin-like fold[1]. The cellulose binding site on the protein surface consists of a linear arrangement of aromatic residues (tyrosines, phenylalanines, and tryptophans)[13], which for *Bs* EXLX1 includes two tryptophans at positions 125 and 126, and a tyrosine at position 157[10]. A further key amino acid residue required for wall extension activity is a lysine at position 119[10]. The C-terminal domain of *Bs* EXLX1 belongs to family 63 of carbohydrate binding modules (CBM63)[10], which mediate binding to polysaccharides[23,24].

In this paper, we analyzed the similarity between "expansin-like proteins" (such as GH45s, loosenins, swollenins, cerato-platanins, EXPNs, and expansin-like proteins found in nematodes) and expansin domains on sequence level by establishing the Expansin Engineering Database (ExED), which collects characterized and putative expansin homologues. The protein sequences in the ExED were divided into different superfamilies ('Bacterial expansins', 'Fungal expansins', and 'Plant expansins') according to sequence identity, and not by phylogenetic relationships of expansins, which were analyzed in[8]. By annotating the two expansin domains and using a continuous standard numbering scheme, conserved sequence motifs of the expansin protein family were identified that could be applied in the screening of genomic data for the identification of novel expansins.

Methods

Sequence collection for the ExED

The Expansin Engineering Database (ExED, https://exed.biocatnet.de) was built within the BioCatNet database system starting from twenty-five protein seed sequences (**Table S1**)[25]. These seed sequences were used as queries for the Basic Local Alignment Search Tool (BLAST)[26] using an e-value cutoff of 10^{-10} against the non-redundant protein database[27] of the National Center for Biotechnology Information (NCBI)[28] and the Protein Data Bank (PDB)[29]. Two subsequent updates were performed to further enrich the ExED. For the first update, the sequences found by the initial search were clustered by UCLUST from the USEARCH package (version 11.0.667)[30] by a threshold of 80% sequence identity, and the centroids (representative sequences) served as seed sequences for a BLAST search in the NCBI non-redundant protein database and the PDB. The seed sequences for the database updates of the ExED are available under https://doi.org/10.18419/darus-622. For the second update, profile hidden Markov models (HMMs) were generated for the N- and C-terminal expansin domains as described below. Further sequences were collected by searching with the *hmmscan* command from the HMMER software package (version 3.1b2, http://www.hmmer.org, Howard Hughes Medical Institute, Chevy Chase, MD, USA)[31]. The hits were filtered by a minimal domain-based score of 35 (chosen after comparison with HMMER's domain-based "in-

dependent" e-values), a minimal hit length of 60 amino acids, and a maximal ratio of bias over domain-based score of 10%.

Sequence hierarchy in the ExED

The initial twenty-five seed sequences comprise six bacterial, one fungal, and seventeen plant expansins, as well as one expansin-like swollenin sequence. The BLAST hits for each of these seed sequences were assigned to a corresponding superfamily named 'Bacterial expansins', 'Fungal expansins', 'Plant expansins', and 'N-terminal domains'. Hence, the division of the identified protein sequences into the different superfamilies was based on sequence identity, and not on phylogenetic relationships. Herein, the term family refers to a group of sequences sharing a certain degree of similarity, i.e. rather a cluster of similar sequences than a clade in a phylogenetic tree. Homologous families were created by a cutoff of 60% pairwise sequence identity as determined by the Needleman-Wunsch algorithm implemented in the EMBOSS software suite (version 6.6.0), with gap opening and extension penalties of 10 and 0.5, respectively[32,33]. All sequence entries which shared at least 98% global sequence identity were assigned to a single protein entry. For each sequence entry, the respective superfamily, homologous family, and protein entry were annotated together with the identifiers of the original source database.

Profile HMMs

A profile hidden Markov model (HMM)[31] was derived for each expansin domain from a multiple sequence alignment built from twenty-eight representative protein sequences, including twenty-two of the twenty-five seed sequences mentioned above, two fungal sequences, and four sequences for which their structure was known (**Table S2**). To determine the region of the two domains in a multiple sequence alignment, four crystal structures of expansins were superimposed (PDB entries 1n10, chain A; 2hcz, chain X; 4fer, chain B; and 4jjo, chain A). The structure-based multiple sequence alignment (**Figure S1**) was generated by the Clustal Omega package[34] (version 1.2.1-1) and STAMP[35] (version 4.4), and visualized by PyMOL[36] (version 4.60, Schrödinger, New York, NY, USA). Based on the structural alignment and on annotations of secondary structures in Pfam[37] (entries PF03330.17 for the N-terminal domain and PF01357.20 for the C-terminal domain), the respective domains were manually retrieved. The individual profile HMMs for the N- and C-terminal expansin domains were built by HMMER from the multiple sequence alignments. The input multiple sequence alignments were aligned against the derived output profile HMMs with the *hmmalign* command from HMMER in order to determine whether there are shifts between the input and output alignments. Shifted alignment columns were refined manually with respect to the positions of known secondary structure elements. The refined profile HMMs of the N- and C-terminal expansin domain comprise 95 and 75 positions, respectively (**Figures S2** and **S3**), and are available together with their underlying alignments at https://doi.org/10.18419/darus-623.

Standard numbering schemes

For the N- and C-terminal expansin domains, standard numbering schemes were introduced to annotate equivalent positions[38]. The *B. subtilis* expansin, *Bs* EXLX1 (PDB entry 4fer), was used as the reference sequence for the assignment of standard position numbers to the sequence entries in the ExED upon alignment against the respective profile HMM and subsequent transfer of position numbers: For both expansin domains, the standard positions range from 11 to 105 and 114 to 186. Insertions with respect to the reference sequence, such as loops, were specified by subsequent decimals. Thus, all position numbers mentioned herein are based on the reference *Bs* EXLX1, unless otherwise stated. Due to insertions in the reference sequence of *Bs* EXLX1, some regions in the underlying multiple sequence alignments of the standard numbering schemes appeared inaccurate, i.e. these regions could not be aligned properly: In the N-terminal expansin domain, inaccurate positions are from 14.1 to 17, 39.1 to 47, and 104.1 to 105; in the C-terminal expansin domain, inaccurate positions are from 162.1 to 164 and 185.1 to 186.

Conservation analyses

The two standard numbering schemes were used to analyze the amino acid frequencies for the two expansin

domains. The domains were annotated by using *hmmscan* against all sequence entries of the ExED and deploying the match criteria mentioned above. Each annotated domain position was analyzed for conserved amino acids. Groups of amino acids with similar biochemical properties, such as charge or polarity, were also taken into account[39,40]. Conservation analyses were performed separately for each superfamily of the ExED, and additionally for EXPA, EXPB, EXLA (expansin-like A), and EXLB (expansin-like B) (**Tables S3** and **S4**). An amino acid position was defined as conserved if it occurred in at least 70% of all annotated sequence entries. Conserved positions were compared with the positions in the structures of two bacterial expansins (PDB entries 4fer, chain B and 4jjo, chain A) and two plant expansins (PDB entries 1n10, chain A and 2hcz, chain X) to predict their functional relevance.

Co-evolution of expansin domains

For comparison of the co-occurrence of the two expansin domains, all sequence entries from the ExED were aligned against the two profile HMMs for the expansin domains. Profile-to-sequence alignments were performed with the *hmmsearch* command from the HMMER software suite with the*max* option to collect all domain-based scores for each possible alignment. The lists of domain-based scores were sorted by sequence identifiers to ensure comparability, and in case of multiple hits, only the maximal bit score was kept. The bivariate histogram was visualized as heat map for bit scores greater than zero in MATLAB (version R2019a, The MathWorks, Natick, MA, USA).

Sequence length distributions

For comparing the lengths of the sequence entries in the ExED, histograms and boxplots were created with MATLAB to visualize frequency distributions and to identify possibly fragmented or artificial sequences (version R2019a, The Mathworks, Natick, MA, USA version 2019a, Statistics and Machine Learning Toolbox version 11.5). The whisker length in a boxplot was chosen as 1.5 times the interquartile range.

Protein sequence networks

Protein sequence networks visualize large sequence datasets as nodes in an undirected graph with edge weights to derive relationships between different clusters or communities. The protein sequences in the ExED were sorted by decreasing sequence length and were subsequently clustered using the USEARCH algorithm (UCLUST) with a threshold of 90% sequence identity (without terminal gaps) to determine a reduced set of centroid sequences (representative sequences)[30]. For each centroid sequence, the N- and the C-terminal expansin domains were annotated by the two profile HMMs with the filter criteria mentioned above. Pairwise sequence identities between two sequences were derived from global Needleman-Wunsch alignments as described above and used as edge weights. Protein sequence networks were generated with edge weights of pairwise sequence identity, filtered by a pre-defined threshold. Metadata of the nodes (e.g. the sequence ID) and of the edges (i.e. the edge weights) were summarized in GraphML files by applying the NetworkX library in Python (version 1.9) for an automated assignment of node and edge attributes [41]. The GraphML files are available at https://doi.org/10.18419/darus-624. Protein sequence networks were visualized with Cytoscape version 3.7.2[42] using a prefuse, force-directed layout with respect to the edge weights.

For the networks showing the relationships between CBM63s and expansin homologues, and between GH45s and the N-terminal expansin domain homologues, CD-HIT (version 4.7) was used with a clustering threshold of 90% and a word size of 5 (instead of UCLUST)[43,44]. The GH45 sequences were downloaded from the protein family database (Pfam, version 32.0, accession PF02015)[45], whereas the CBM63 sequences were downloaded from the carbohydrate-active enzymes (CAZy) database on June 3, 2019[46]. In the CAZy database, 633 individual CBM63 sequences were deposited, but only 582 NCBI accessions were available at the time of writing, as some of the records were moved or entries were merged. Members of CBM63 were annotated by the profile HMMs for the two expansin domains (https://doi.org/10.18419/darus-625).

Homologous expansin-like domains in other proteins

In order to find similarities between the expansin family and the GH45 endoglucanase family, a structure-

based multiple sequence alignment was performed of the N-terminal expansin domains of five expansins (PDB entries 2bh0, 4jjo, 4fer, 1n10, and 2hcz) and the complete sequences of seven GH45s (PDB entries 1eng, 1hd5, 1oa9, 1wc2, 5h4u, 5kjo, and 5xbu). Five representative sequences of expansin-like proteins were analyzed: swollenin from *Trichoderma reesei* (NCBI accession AJ245918.1)[47], loosenin from *Bjerkandera adusta*(NCBI accession ADI72050.2)[48], EXPN from*Endogone* sp. FLAS-F59071 (NCBI accession RUS20349.1)[49,50], an expansin-like protein found in nematode *Heterodera glycines* (NCBI accession ADL29728.1)[51], and cerato-platanin from*Ceratocystis platani* (NCBI accession CAC84090.2)[52]. The two profile HMMs of the N- and C-terminal expansin domains were used to search within these five expansin-like protein sequences for expansin domains with the filter criteria mentioned above.

Identification of expansin domains in actinobacterial genomes

Five actinobacterial genomes were selected to show the application of the ExED for the identification of expansin domains. An Illumina MiSeq sequencer was used to sequence the genomes (NGS facility, University of the Western Cape, South Africa). Due to the high G+C content of actinobacterial DNA, a 10% PhiX spike was included in the run. The genomes were assembled using the A5-miseq pipeline[53].

The two newly created profile HMMs mentioned above were applied to search the five actinobacterial genomes for the occurrence of expansin domains. Nucleic acid sequences were translated using the default codon usage table available in the *transeq* tool from the EMBOSS software suite (version 6.6.0[54]). Translated amino acid sequences with less than 60 subsequent amino acid symbols were discarded to reduce computation time.

The *hmmscan* tool from the HMMER software suite (version 3.1b2, http://www.hmmer.org, Howard Hughes Medical Institute, Chevy Chase, MD, USA) was used to scan the translated amino acid sequences with profile HMMs. The hits from *hmmscan* were filtered by a minimal domain-based score of 35 and a minimal coverage of 75% (defined as the ratio of hit length without insertions divided by the length of the profile HMM).

The matches for the profile HMMs of expansin domains were extended to find the adjacent start methionine and stop codon along the contig sequence of each match. The first or last available amino acid position in a contig was used to extend the hits, in case of a missing start or stop codon, respectively. The extended hit sequences are available for download under https://doi.org/10.18419/darus-699.

Results

The Expansin Engineering Database (ExED)

The current version of the ExED contains 15,089 sequence entries, 12,400 protein entries, and twenty-one protein structures (**Tables 1**and **S5**), which, based on global sequence similarity, were assigned to four superfamilies (comprising 12,404 sequence entries, 9954 protein entries and seventeen structures). Three superfamilies include expansin homologues with two domains and were named according to their dominant source organisms: superfamily 1 'Bacterial expansins' (1172 sequences, ten structures), superfamily 2 'Fungal expansins' (543 sequences, no structure), and superfamily 3 'Plant expansins' (8269 sequences, six structures). The members of superfamily 4 'N-terminal domains' consist of the N-terminal expansin domain only (2420 sequences, one structure). This superfamily comprises eukaryotic and bacterial sequences, e.g. from *Magnoliophyta* (A, B, and C),*Actinobacteria* , *Oomycetes* , and *Basidiomycota* . The remaining number of 2685 sequences (corresponding to 2446 protein entries) and four structures could not be assigned to the four superfamilies and was thus collected in an unclassified fifth superfamily, which was omitted for further investigations.

The sequence lengths in the superfamilies 'Bacterial expansins', 'Fungal expansins', and 'Plant expansins' vary between 40 and 1400 amino acids with a sharp peak between 250 and 270 amino acids and two minor peaks at 150 and at 600 amino acids (**Figure S4**). The sequence length distributions differ for each of the four superfamilies (**FigureS5**). For further analysis of whole expansin sequences and comparison with expansin-like proteins, only sequences with a length between 210 and 300 amino acids were considered (7706 sequences from the superfamilies 'Bacterial expansins', 'Fungal expansins', and 'Plant expansins') (**Figure S4**).

In the protein sequence network, which was built from global sequence alignments for the superfamilies 'Bacterial expansins', 'Fungal expansins', and 'Plant expansins', the latter are the most frequent group forming four large separate clusters, which consist of one or more homologous families (Hfams): cluster A has been classified as EXPA (Hfams 9-20), cluster B as EXPB (Hfams 21, 22), cluster C as EXLB (Hfams 24, 25), and cluster D as EXLA (Hfam 23) (**Figure 2**). These four clusters are followed by two clusters of the superfamily 'Fungal expansins' (Hfam 7) and three clusters of the superfamily 'Bacterial expansins' (Hfams 3, 4; 1, 2, 4, 6; and 3). Noteworthy, the 'Plant expansins' clusters A and D also contain bacterial sequences.

In our study, expansins were found in *Bacteria* , *Archaea* , and *Eukaryota* . When looking in detail at the major taxa in the tree of life (after Fig. 1 in[8]), expansins occur in *Gammaproteobacteria* ,*Betaproteobacteria* , *Deltaproteobacteria* ,*Acidobacteria* , *Bacteroidetes* , *Fibrobacteres* ,*Ignavibacteria* , *Actinobacteria* , *Chloroflexi* ,*Firmicutes* , *Cyanobacteria* (all *Bacteria*),*Euryarchaeota* (*Archaea*), *Metazoa* , Fungi,*Evosea* , *Discosea* , *Discoba* , *Embryophyta* ,*Chloroplastida* , *Rhodophyta* , and *Stramenopiles* (all*Eukaryota*) (**Table S6**).

Sequence space of expansin domains

Two profile HMMs for the N-terminal and the C-terminal expansin domains were derived and used for annotation of the two domains in all 12,404 classified sequences of the ExED (superfamilies 1, 2, 3, and 4), independent of their sequence lengths. For the superfamilies 'Bacterial expansins', 'Fungal expansins', and 'Plant expansins', the N- and the C-terminal expansin domains could be annotated in 9,470 out of 9,984 sequences and in 8,896 out of 9,984 sequences, respectively (**Table S5**). In 2,182 out of the 2,420 sequences from the superfamily 'N-terminal domains', only the N-terminal expansin domain was annotated.

Based on the annotated domains in the classified superfamilies, two protein sequence networks were generated. The sequence network of N-terminal expansin domains is dominated by three large clusters (**Figure S6**): Homologues of cluster A classified as EXPA (Hfam 9-20), homologues of cluster B as EXPB (Hfam 21, 22), and homologues of cluster C as EXLX as well as fungal sequences (Hfam 3, 4, 8). These clusters are supplemented by clusters D (Hfam 24, 25; EXLB), E (Hfam 26;*Magnoliophyta* A), F (Hfam 23; EXLA), G (Hfam 7; Fungi), H (Hfam 27; *Magnoliophyta* B), and cluster I comprising N-terminal domains from different sources (Hfam 8, 11, 31, 32; Fungi, EXPA,*Basidiomycota* , Loosenin). The N-terminal domains of*Magnoliophyta* B, *Actinobacteria* , and *Oomycetes* form separate clusters. The sequences of CBM63 are within clusters of homologous families 3 and 4 from the superfamily 'Bacterial sequences'.

The sequence network of the C-terminal expansin domain is dominated by six large clusters from 'Plant expansins', previously annotated as EXPA, EXPB, EXLB, and EXLA (clusters A-C and E-G), one cluster from 'Fungal expansins' (D, Hfam 7), and three clusters from 'Bacterial expansins' (H-J, Hfams 1, 3, 4, 6) (**Figure S7**). In each of the two domain-based networks, one bacterial sequence was found in a cluster from 'Plant expansins', *Streptomyces acidiscabies* (NCBI accession GAQ55178.1) in EXPA (**Figure S6**), and *Soehngenia saccharolytica* (NCBI accession TJX44964.1) in EXLA (**Figure S7**).

The N- and C-terminal expansin domains have not evolved independently, but have co-evolved, as indicated by the correlation of sequence similarities of the two domains to the respective profile HMM (**Figure 3**). The shift in respect to the diagonal indicates a higher conservation for the N-terminal expansin domain than for the C-terminal expansin domain.

Conserved positions in the two expansin domains

Standard numbering schemes for the N- and the C-terminal expansin domains (from positions 11 to 105 and 114 to 186, respectively) were applied to identify conserved positions ([?]70% occurrence) in the superfamilies 'Bacterial expansins', 'Fungal expansins', and 'Plant expansins' (**Table 2**). In both expansin domains, glycine was the most frequently conserved amino acid (**Tables S3** and**S4**). In the N-terminal expansin domain, nine positions were conserved in the three superfamilies ('Bacterial expansins', 'Fungal expansins', and 'Plant expansins'): threonine 12, glycine 21, alanine 36, glycine 53, proline 74, aspartate 82, leucine

83, phenylalanine 88, and glycine 97 (**Table S3**). In the superfamily 'Bacterial expansins', further seven positions were highly conserved ([?]90% occurrence): valine 58, glycine 60, glycine 63, aspartate 71, serine 84, alanine 87, and isoleucine 91; in the superfamily 'Fungal expansins', additional highly conserved amino acids are cysteine 23, phenylalanine 25, tryptophan 44, cysteine 52, cysteine 55, methionine 68, and leucine 81; and in the superfamily 'Plant expansins' highly conserved amino acids are alanine 22, cysteine 23, glycine 24, glycine 49, cysteine 52, cysteine 55, cysteine 60, cysteine 61.8, threonine 70, and phenylalanine 81.

Due to our conservation analysis, five of the previously proposed six cysteines[14] were highly conserved in the superfamily 'Plant expansins', three conserved cysteines were found in the superfamily 'Fungal expansins', and none in the superfamily 'Bacterial expansins' (**Table 2** and https://doi.org/10.18419/darus-735). The conserved cysteines at standard positions C23 and C52, C55 and a cysteine upstream of the N-terminal expansin domain standard numbering, and C60 and 61.8 were proposed to form disulfide bonds[14].

In the C-terminal expansin domain, only two positions are conserved in the three superfamilies 'Bacterial expansins', 'Fungal expansins', and 'Plant expansins': tryptophan 149 and glycine 179 (**Table S4**). In addition, in the superfamily 'Bacterial expansins', highly conserved positions are lysine 119, glycine 121, tryptophan 126, proline 137, tyrosine 157, asparagine 158, glycine 166, threonine 175, and aspartate 176; in the superfamily 'Fungal expansins', glycine 121, serine 123, tryptophan 126, phenylalanine 127, glutamine 130, valine 131, asparagine 133, valine 143, serine 143.1, aspartate 146, arginine 154, tyrosine 157, asparagine 158, phenylalanine 160, glycine 164, valine 172, and threonine 175; and in the superfamily 'Plant expansins' highly conserved amino acids are glycine 136, glycine 157, and tryptophan 160. From the conserved positions, superfamily-specific motifs were derived. In the N-terminal domain of 'Bacterial expansins', these motifs are VpGP (standard positions 58-61, "p" is the abbreviation for polar amino acids selected from [40]) and HLDL (80-83) (**Table 3**); in the superfamily 'Fungal expansins' the motifs T(F/W)YG (12-14 and 14.1), GTAnS (34-38, "n" is the abbreviation for non-polar amino acids selected from[40]), VpGn (58-61), and HLDL (80-83) were identified; and in the superfamily 'Plant expansins', the motifs T(F/W)YG (12-14 and 14.1) in EXPA and EXPB, GGACGYG (20-26) in minor modifications in all four plant expansin groups, and HFDL (80-83) in EXPA and EXPB were identified (**Tables 3** , **S3** and https://doi.org/10.18419/darus-735). In the C-terminal expansin domain of the superfamily 'Bacterial expansins', the motif QVRNH (130-134) was conserved; in the superfamily 'Fungal expansins' the motifs QVnN (130-133), LEVSTDGD (141-146, including 143.1 and 143.2 as insertions relative to the reference sequence), GGG (164-166), and VDVRVT (170-175) were identified. At the standard positions 170 to 175, the motif LSFpVT is included in sequences of EXPA. Sequences from the superfamilies 'Bacterial expansins' and 'Fungal expansins' were found to share the motif, KpG(S/T)S (119-123); and pGS exists also in EXPB. EXLA and EXLB were found to share the motif, YLA (126-128). The motif WGA exists in minor modifications in all four groups of 'Plant expansins' (**Tables 3** , **S4** and https://doi.org/10.18419/darus-735).

Homologous expansin-like domains in other proteins

The GH45 protein sequences show conserved positions, which are also highly conserved in the superfamilies 'Plant expansins' and 'Fungal expansins': the EXPA/EXPB motif HFDL (80-83), glycine 21, cysteine 23 and glycine 24 of the plant motif GGACGYG (21-26), threonine 12 and tyrosine 14 of the plant and fungal motif T(F/W)YG (12-14 and 14.1), and alanine 36 of the fungal motif GTAnS (34-38) (**Figure S8**). Thus, on a local sequence level, GH45 endoglucanases are more similar to the N-terminal expansin domain than expected from their different global protein sequences (**Figure 4**).

For further comparison, 582 protein sequences of the carbohydrate-binding module family 63 (CBM63) with a sequence length between 57 and 746 amino acids were downloaded from the CAZy database[46]. Interestingly, 511 of these sequences contained both expansin domains and were therefore already annotated in the ExED in the superfamilies 'Bacterial expansins' and 'Fungal expansins'. Four CBM63 sequences contained only the C-terminal expansin domain, whereas 58 CBM63 sequences contained only the N-terminal expansin domain and shared a sequence identity of over 60% with N-terminal expansin domains of the superfamily 'Bacterial expansins' (**Figure S6**). A protein sequence network including the whole CBM63 sequences and expansin sequences from the superfamilies 'Bacterial expansins', 'Fungal expansins', and 'Plant expansins' revealed

the similarity of CBM63 sequences to 'Bacterial expansins' and also to 'Fungal expansins' from homologous family 7 (**Figure 5)** .

The members of the superfamily 'N-terminal domains' consist of the N-terminal expansin domain only. Similarly, loosenin (NCBI ADI72050.2), EXPN from *Endogone* sp. FLAS-F59071 (NCBI accession RUS20349.1), the expansin-like protein found in nematode*Heterodera glycines* (NCBI ADL29728.1), and cerato-platanin from *Ceratocystis platani* (NCBI accession CAC84090.2) consist only of the N-terminal expansin domain (**Table S7**). At a threshold of 60% sequence identity, the N-terminal domains of loosenin and Basidiomycota cluster with fungal sequences from Hfam 7 and plant sequences from Hfam 11 (**Figure S6**). In contrast, swollenin was found to possess only a distantly related C-terminal expansin domain (**Table S7**).

Annotation of expansin domains in actinobacterial genomes

As a case study for the application of ExED in genome sequence annotation, actinobacterial genomes from various South African habitats were analyzed for the presence of expansin domains and conserved amino acid positions, using the profile HMMs of the expansin domains (**Tables S8** and **S9**). In general, the sequence regions identified for the N-terminal expansin domains emerged with higher HMMER scores, whereas the C-terminal domains seemed less conserved (compare with **Figure 3**). Despite the lower scores for the C-terminal expansin domain, the coverage for the underlying profile HMM was still high (90%). One genome hit was identified in sediment samples collected at Gamka River in the Swartberg Mountain Range, which was identical to an expansin homologue from *Streptomyces swartbergensis* (NCBI accession WP_086602418), which matched well the profile HMM of the N-terminal expansin domain (score: 60, 98% coverage) and moderately the profile HMM of the C-terminal expansin domain (score: 19, 89% coverage). The sequence from *S. swartbergensis* contains amino acids that are conserved in the superfamily 'Bacterial expansins' (threonine 12, glycine 21, alanine 36, glycine 53, tyrosine 55, proline 74, aspartate 82, leucine 83, phenylalanine 88, and glycine 97 in the N-terminal expansin domain; lysine 119, tryptophan 126, tryptophan 149, tyrosine 157, and glycine 179 in the C-terminal expansin domain) and also amino acids that are conserved in the superfamilies 'Fungal expansins' or 'Plant expansins' (tyrosine 14, cysteine 23, cysteine 52, and cysteine 73).

Discussion

Expansins typically consist of about 225 amino acids (about 26 kDa) and an N-terminal signal peptide[2], in total 250 to 275 amino acids[55], which is in agreement with the average sequence length of 262 amino acids identified in this study. Thus, sequences shorter than 210 amino acids or longer than 300 amino acids were excluded from global sequence analyses (**Figure S4**). However, sequences with a length of about 600 amino acids contained replications of expansin domains as fusion proteins or due to sequencing errors, leading to expansin sequences that contained each domain two or three times. Since the two expansin domains have a length between 80 and 90 amino acids, shorter protein sequences can be considered as fragments or incomplete expansin domains.

The occurrence of expansins in major taxa in the tree of life (after Fig. 1 in[8] where a comprehensive phylogenetic analysis of expansin genes across all kingdoms of life is shown) is comparable to the results obtained in this study (https://doi.org/10.18419/darus-693). For twelve out of ninety groups that were compared, the results are different, e.g. the archaeon *Halomicroarcula* sp. LR21 can be found in the ExED and contains one expansin homologue for which both expansin domains are annotated, whereas previous studies in[8] did not find a putative expansin in *Archaea* . Other, apparently hitherto unknown, occurrences of putative expansins in the ExED include thirty-six sequences of *Fibrobacteres* , one sequence of *Ignavibacteria* in which both expansin domains can be found, seven sequences of *Discosea* , one sequence of*Discoba* , and one sequence of *Acidobacteria* , but without domain annotations. Further expansin sequences that were not included in this study but mentioned in [8] are from the taxa*Verrucomicrobia* , *Chlorobi* , *Tubulinea* ,*Glaucophyta* , *Haptophyta* , *Dinoflagellata* , and*Phaeophyta* .

The protein sequence networks confirmed the nomenclature and classification of expansins into three kingdoms of *Bacteria* , Fungi, and *Viridiplantae* and the subclassification of plant expansins into EXPA, EXPB, EXLA, and EXLB[56](**Figure 2**). Despite the differences on global sequence level, the protein sequence

networks of expansins from *Bacteria* and *Viridiplantae* share similarities on a domain-based sequence level (**Figures S1, S6** and **S7**). The N-terminal expansin domain is more conserved than the C-terminal expansin domain (**Figure 3** and https://doi.org/10.18419/darus-735). When expansin homologues from more diverse backgrounds are discovered in the future, updated profile HMMs will show more insights into the possible co-evolution of both expansin domains.

A conservation analysis revealed and confirmed positions with an essential functional or structural role in expansin homologues. Glycine is structurally relevant, as it mediates the formation of short loops[57] and is frequently observed at the N- and C-caps of α-helices to increase helix stability[58]. As observed previously for other protein families[59,60], glycine is the most conserved amino acid in both expansin domains. In expansins, all four conserved glycines are located in loop regions (**Table 2** , compare with **Figure 1**). The conservation of threonine 12 and aspartate 82 in *Bacteria* , Fungi, EXPA, and EXPB confirms their functional role[10]. Interestingly, at standard position 75, which plays a moderate role in cell wall extension activity of *Bs* EXLX1[10], a glutamate is conserved in the superfamily 'Bacterial expansins', and a glycine in EXPA and EXLB. In contrast, standard position 75 is not conserved in the superfamily 'Fungal expansins', in EXPB, and in EXLA (**Table 2**and https://doi.org/10.18419/darus-735). Aspartate 71, which has been proposed as important but not essential for wall extension activity of*Bs* EXLX1[10], is conserved in the superfamilies 'Bacterial expansins' and 'Fungal expansins', and in EXPB, EXLA, and EXLB (**Table 2** and https://doi.org/10.18419/darus-735). However, three other proposed key amino acids for cell wall extension activity (threonine 14, serine 16, and tyrosine 73[10]) are neither conserved in expansins from *Bacteria* , Fungi, nor *Viridiplantae* (**Table S3** and https://doi.org/10.18419/darus-735), indicating the importance of an increased sample size for conservation analysis. The large number of expansin sequences investigated here also provided a deeper insight into the structural or functional relevance of disulfide bridges in the different superfamilies. Previously, three disulfide bridges were proposed to stabilize the tertiary structure of the N-terminal expansin domain of EXPA and EXPB[14,15]. Five of the proposed six cysteines could be confirmed as highly conserved in the superfamily 'Plant expansins' (**Table 2** and https://doi.org/10.18419/darus-735). The sixth cysteine is located directly before the linker to the C-terminal expansin domain and therefore not included in our profile HMM for the N-terminal expansin domain. Against expectations, the additional highly conserved forth cysteine pair in plant α-expansins from [15] was not found in our analysis(https://doi.org/10.18419/darus-735) . Only three conserved cysteines were found in the superfamily 'Fungal expansins', thus not all fungal expansin homologues possess three disulfide bridges, as concluded from the expansin *Sc* Exlx1[16]. None of the six cysteines was conserved in the superfamily 'Bacterial expansins' (**Table 2**), which is in accordance with previous observations of bacterial expansins lacking disulfide bridges[13].

In the C-terminal expansin domain, the three aromatic residues at standard positions 125, 126, and 157, which mediate binding to cellulose[10], are conserved in the superfamilies 'Bacterial expansins' and 'Fungal expansins', but are less conserved in the superfamily 'Plant expansins' (**Table S4** and https://doi.org/10.18419/darus-735). Lysine 119, which is important for cell wall-loosening activity[10], is conserved in the superfamilies 'Bacterial expansins' and 'Fungal expansins', but not conserved in the superfamily 'Plant expansins' (**Table 2**).

Through the use of conservation analysis, previously published family-specific motifs were confirmed: in the N-terminal expansin domain, the T(F/W)YG motif was present in the two superfamilies 'Bacterial expansins' and 'Fungal expansins' (standard positions 12-14 and 14.1), and the motifs GGACG (20-24) and HFD (80-82) in the superfamily 'Plant expansins'[9,55] (**Table 3**). We suggest to extend the GGACG motif to a GGACGYG motif and the HFD motif to a HFDL motif in plant expansins. In bacterial and fungal expansins, these two plant motifs are slightly different: in the superfamily 'Fungal expansins', the GGACGYG motif is shorter (GGxC), and in fungal and bacterial expansins the HFDL motif is replaced by HLDL. The HLD motif as well as the GGACS motif were already described for the fungal expansin *Sc* EXLX1[16]. Newly proposed motifs in the N-terminal expansin domain are VpGP (58-61) in the superfamily 'Bacterial expansins' and GTAnS (34-38) in the superfamily 'Fungal expansins' (**Tables 3** and **S3)** , where p and n denote polar and nonpolar amino acids, respectively. In expansins from Fungi, the proline of the VpGP-motif is replaced by a non-polar amino acid. The previously described CDRC-motif at the amino

terminus of EXLA[55] is located beyond the boundaries of our profile HMM for the N-terminal expansin domain .

No sequence motifs have been proposed yet for the C-terminal expansin domain, whereas we found eight novel motifs: KpG(S/T)S (119-123) and QVRNH (130-134) in the superfamilies 'Bacterial expansins' and 'Fungal expansins', where QVRNH shows slight modifications; LEVSTDGD (141-146, including 143.1 and 143.2), GGG (164-166), and VDVRVT (170-175) in 'Fungal expansins'; YLA (126-128) in EXLA and EXLB; WGA (156-158) in EXPB and with slight modifications in EXPA, EXLA and EXLB; and LSFpVT (170-175) in EXPA (**Table 3**). When annotating expansin sequences in future studies, these sequence motifs will help to assign unknown protein sequences (e.g. metagenomic sequences) to the kingdoms *Viridiplantae* , *Bacteria* , or Fungi, and to distinguish the plant expansins EXPA, EXPB, EXLA, and EXLB (**Tables 3** ,**S3** , and **S4**). Exemplary annotations were shown herein for actinobacterial genome samples from South Africa, including a putative expansin homologue from *S. swartbergensis*(**Tables S8** and **S9**).

The large number of expansin sequences used for analysis not only improved the identification of motifs, but also shed light on evolutionary relationships. Interestingly, when searching with the newly established profile HMMs for expansin domains within the CBM63 protein sequences from CAZy, 510 out of the 582 CBM63 protein sequences were found to contain both expansin domains (**Table S10**). Only four sequences had a similarity to the C-terminal expansin domain, while missing the N-terminal expansin domain, as suggested previously[1], and 58 CBM63 sequences contained only the N-terminal expansin domain.

The observation of four bacterial sequences being found in clusters of plant expansins supports the hypothesis that microbial expansins were derived via horizontal gene transfer from plants to microbes[7] (**Figures 2** , **S6** , and**S7**). The two bacterial sequences in clusters of the superfamily 'Plant expansins' (**Figure 5**) are from the plant pathogens*Kutzneria* sp. 744 (NCBI accession EWM10128.1) and*Streptomyces acidiscabies* (NCBI accession WP 050370046.1), which are both actinobacteria, as described previously[2].

With the chosen filter criteria, the sequence of the fungal swollenin does not contain any expansin domain. As the score for the C-terminal expansin domain is far below the chosen criteria, the swollenin sequence resembles a distantly related C-terminal expansin domain (**Table S7**), but we found no N-terminal expansin domain within the protein sequence of swollenin. This is due to the short N-terminal expansin domain in the swollenin from *Trichoderma reesei* and confirms the rather low sequence similarity between swollenin and expansins[47].

On a global sequence level, GH45s and N-terminal expansin domains share less than 30% pairwise sequence identity (**Figure 4**), and neither the profile HMM search of the N- and C-terminal expansin domains in the 542 GH45 sequences nor the profile HMM search of the GH45 profile HMM from Pfam (https://pfam.xfam.org/family/PF02015/hmm) in the 15,089 sequences of the ExED resulted in a match. In comparison to N-terminal expansin domains, GH45 sequences are longer due to several inserts and longer loop regions (179-208 amino acids as compared to 90-115 amino acids of the N-terminal expansin domains). Despite these differences, the evolutionary relationship between the two protein families is underlined by conserved amino acids. Both the conserved threonine and aspartate at standard positions 12 and 82, and the HFDL-motif (standard positions 80-83) were found in the GH45 protein sequences.

This study confirms the observation that microbial expansins comprise two protein domains and are widely distributed across diverse lineages of *Archaea* ,*Bacteria* , Fungi, other eukaryotic microbes[8], and *Viridiplantae* . Therefore, the ExED can serve as a basis for a more detailed phylogenetic analysis in order to elucidate the origin of expansins and ancient evolutionary dynamics. Furthermore, the ExED can be used to search for expansin genes in virulent fungal and bacterial plant pathogens.

References

1. Cosgrove DJ. Microbial Expansins. *Annu Rev Microbiol* . 2017;71(1):479-497. doi:10.1146/annurev-micro-090816-093315

2. Georgelis N, Nikolaidis N, Cosgrove DJ. Bacterial expansins and related proteins from the world of microbes. *Appl Microbiol Biotechnol* . 2015;99(9):3807-3823. doi:10.1007/s00253-015-6534-0

3. Cosgrove DJ. Plant expansins: Diversity and interactions with plant cell walls. *Curr Opin Plant Biol* . 2015;25. doi:10.1016/j.pbi.2015.05.014

4. Cosgrove DJ. Catalysts of plant cell wall loosening.*F1000Research* . 2016;5:119. doi:10.12688/f1000research.7180.1

5. Tovar-Herrera OE, Rodríguez M, Olarte-Lozano M, et al. Analysis of the Binding of Expansin Exl1, from Pectobacterium carotovorum, to Plant Xylem and Comparison to EXLX1 from Bacillus subtilis. *ACS Omega* . 2018;3(6):7008-7018. doi:10.1021/acsomega.8b00406

6. McQueen-Mason S, Durachko DM, Cosgrove DJ. Two endogenous proteins that induce cell wall extension in plants. *Plant Cell* . 1992;4:1425-1433. doi:10.2307/3869513

7. Nikolaidis N, Doran N, Cosgrove DJ. Plant Expansins in Bacteria and Fungi: Evolution by Horizontal Gene Transfer and Independent Domain Fusion. *Mol Biol Evol* . 2014;31(2):376-386. doi:10.1093/molbev/mst206

8. Chase WR, Zhaxybayeva O, Rocha J, Cosgrove DJ, Shapiro LR. Global cellulose biomass, horizontal gene transfers and domain fusions drive microbial expansin evolution.*New Phytol* . 2020. doi:10.1111/nph.16428

9. Cosgrove DJ. New genes and new biological roles for expansins. *Curr Opin Plant Biol* . 2000;3:73-78. doi:10.1016/S1369-5266(99)00039-4

10. Georgelis N, Tabuchi A, Nikolaidis N, Cosgrove DJ. Structure-function analysis of the bacterial expansin EXLX1. *J Biol Chem* . 2011;286:16814-16823. doi:10.1074/jbc.M111.225037

11. Georgelis N, Yennawar NH, Cosgrove DJ. Structural basis for entropy-driven cellulose binding by a type-A cellulose-binding module (CBM) and bacterial expansin. *Proc Natl Acad Sci* . 2012. doi:10.1073/pnas.1213200109

12. Wang T, Chen Y, Tabuchi A, Cosgrove DJ, Hong M. The target of β-expansin EXPB1 in maize cell walls from binding and solid-state NMR studies. *Plant Physiol* . 2016;172:2107-2119. doi:10.1104/pp.16.01311

13. Kerff F, Amoroso A, Herman R, et al. Crystal structure and activity of Bacillus subtilis YoaJ (EXLX1), a bacterial expansin that promotes root colonization. *Proc Natl Acad Sci* . 2008;105:16876-16881. doi:10.1073/pnas.0809382105

14. Yennawar NH, Li AC, Dudzinski DM, Tabuchi A, Cosgrove DJ. Crystal structure and activities of EXPB1 (Zea m 1), a β-expansin and group-1 pollen allergen from maize. *Proc Natl Acad Sci U S A* . 2006;103:14664-14671. doi:10.1073/pnas.0605979103

15. Gaete-Eastman C, Morales-Quintana L, Herrera R, Moya-León MA. In-silico analysis of the structure and binding site features of an α-expansin protein from mountain papaya fruit (VpEXPA2), through molecular modeling, docking, and dynamics simulation studies. *J Mol Model* . 2015;21(5):115. doi:10.1007/s00894-015-2656-7

16. Tovar-Herrera OE, Batista-García RA, Sánchez-Carbente MDR, Iracheta-Cárdenas MM, Arévalo-Niño K, Folch-Mallol JL. A novel expansin protein from the white-rot fungus Schizophyllum commune. *PLoS One* . 2015;10(3):1-17. doi:10.1371/journal.pone.0122296

17. Castillo RM, Mizuguchi K, Dhanaraj V, Albert A, Blundell TL, Murzin AG. A six-stranded double-psi β barrel is shared by several protein superfamilies. *Structure* . 1999;7:227-236. doi:10.1016/S0969-2126(99)80028-8

18. Nomura T, Iwase H, Saka N, Takahashi N, Mikami B, Mizutani K. High-resolution crystal structures of the glycoside hydrolase family 45 endoglucanase EG27II from the snail Ampullaria crossean . *Acta Crystallogr Sect D Struct Biol* . 2019;75(4):426-436. doi:10.1107/s2059798319003000

19. Davies GJ, Dodson GG, Hubbard RE, et al. Structure and function of endoglucanase V. *Nature* . 1993;365:362-364. doi:10.1038/365362a0

20. Suzuki H, Vuong T V., Gong Y, et al. Sequence diversity and gene expression analyses of expansin-related proteins in the white-rot basidiomycete, Phanerochaete carnosa. *Fungal Genet Biol* . 2014;72:115-123. doi:10.1016/j.fgb.2014.05.008

21. Baccelli I, Luti S, Bernardi R, Scala A, Pazzagli L. Cerato-platanin shows expansin-like activity on cellulosic materials. *Appl Microbiol Biotechnol* . 2014;98:175-184. doi:10.1007/s00253-013-4822-0

22. Gourlay K, Hu J, Arantes V, Penttilä M, Saddler JN. The use of carbohydrate binding modules (CBMs) to monitor changes in fragmentation and cellulose fiber surface morphology during cellulase- And swollenin-induced deconstruction of lignocellulosic substrates. *J Biol Chem* . 2015;290(5):2938-2945. doi:10.1074/jbc.M114.627604

23. Tomme P, Van Tilbeurgh H, Pettersson G, et al. Studies of the cellulolytic system of Trichoderma reesei QM 9414: Analysis of domain function in two cellobiohydrolases by limited proteolysis. *Eur J Biochem* . 1988;170(3):575-581. doi:10.1111/j.1432-1033.1988.tb13736.x

24. Gilkes NR, Warren RA, Miller RC, Kilburn DG. Precise excision of the cellulose binding domains from two Cellulomonas fimi cellulases by a homologous protease and the effect on catalysis. *J Biol Chem* . 1988;263(21):10401-10407.

25. Buchholz PCF, Vogel C, Reusch W, et al. BioCatNet: A Database System for the Integration of Enzyme Sequences and Biocatalytic Experiments.*ChemBioChem* . 2016;17:2093-2098. doi:10.1002/cbic.201600462

26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* . 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2

27. Benson DA, Cavanaugh M, Clark K, et al. GenBank. *Nucleic Acids Res* . 2013;41:D36-D42. doi:10.1093/nar/gks1195

28. Federhen S. The NCBI Taxonomy database. *Nucleic Acids Res* . 2012;40:D136-D143. doi:10.1093/nar/gkr1178

29. Berman HM, Kleywegt GJ, Nakamura H, Markley JL. The Protein Data Bank archive as an open data resource. *J Comput Aided Mol Des* . 2014;28:1009-1014. doi:10.1007/s10822-014-9770-y

30. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* . 2010;26:2460-2461. doi:10.1093/bioinformatics/btq461

31. Eddy SR. HMMER: Profile hidden Markov models for biological sequence analysis. *HMMER User's Guid* . 2001. doi:10.1109/TIA.2013.2279901

32. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* . 1970;48:443-453. doi:10.1016/0022-2836(70)90057-4

33. Rice P, Longden L, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* . 2000;16:276-277. doi:10.1016/S0168-9525(00)02024-2

34. Sievers F, Higgins DG. Clustal Omega. *Curr Protoc Bioinforma* . 2014;48. doi:10.1002/0471250953.bi0313s48

35. Russell RB, Barton GJ. Multiple protein sequence alignment from tertiary structure comparison: Assignment of global and residue confidence levels. *Proteins Struct Funct Bioinforma* . 1992;14:309-323. doi:10.1002/prot.34014021

36. DeLano W. Pymol: An open-source molecular graphics tool. *CCP4 Newsl Protein Crystallogr* . 2002:44-53.

37. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer ELL. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res* . 1999;27:260-262. doi:10.1093/nar/27.1.260

38. Vogel C, Widmann M, Pohl M, Pleiss J. A standard numbering scheme for thiamine diphosphate-dependent decarboxylases. *BMC Biochem* . 2012;13. doi:10.1186/1471-2091-13-24

39. Betts MJ, Russell RB. Amino-Acid Properties and Consequences of Substitutions. In:*Bioinformatics for Geneticists: A Bioinformatics Primer for the Analysis of Genetic Data: Second Edition* . 2007. doi:10.1002/9780470059180.ch1

40. Volkenstein M V. Coding of polar and non-polar amino-acids. *Nature* . 1965;207:294-295. doi:10.1038/207294a0

41. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using NetworkX. *7th Python Sci Conf (SciPy 2008)* . 2008:11-15.

42. Shannon P, Markiel A, Owen Ozier 2, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* . 2003;(13):2498-2504. doi:10.1101/gr.1239303.metabolite

43. Fu L, Niu B, Zhu Z, et al. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* . 2014;23:1312-1313. doi:10.1093/bioinformatics/bts565

44. Li W, Godzik A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* . 2006;22:1658-1659. doi:10.1093/bioinformatics/btl158

45. Finn RD, Mistry J, Tate J, et al. The Pfam protein families database. *Nucleic Acids Res* . 2009;38:D211-D222. doi:10.1093/nar/gkp985

46. Cantarel BI, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucleic Acids Res* . 2009;37:D233-238. doi:10.1093/nar/gkn663

47. Saloheimo M, Paloheimo M, Hakola S, et al. Swollenin, a Trichoderma reesei protein with sequence similarity to the plant expansins, exhibits disruption activity on cellulosic materials. *Eur J Biochem* . 2002;269(17):4202-4211. doi:10.1046/j.1432-1033.2002.03095.x

48. Quiroz-Castañeda RE, Martínez-Anaya C, Cuervo-Soto LI, Segovia L, Folch-Mallol JL. Loosenin, a novel protein with cellulose-disrupting activity from Bjerkandera adusta. *Microb Cell Fact* . 2011;10:8. doi:10.1186/1475-2859-10-8

49. Chang Y, Desirò A, Na H, et al. Phylogenomics of Endogonaceae and evolution of mycorrhizas within Mucoromycota. *New Phytol* . 2019. doi:10.1111/nph.15613

50. Varga T, Krizsán K, Földi C, et al. Megaphylogeny resolves global patterns of mushroom evolution. *Nat Ecol Evol* . 2019;222. doi:10.1038/s41559-019-0834-1

51. Zhang YD, Kong XC, Huang WK, et al. Identification and functional analysis of two expansin genes Hg-exp-1 and Hg-exp-2 from the soybean cyst nematode (Heterodera glycines). *Sci Agric Sin* . 2018. doi:10.3864/j.issn.0578-1752.2018.17.006

52. Pazzagli L, Cappugi G, Manao G, Camici G, Santini A, Scala A. Purification, characterization, and amino acid sequence of cerato- platanin, a new phytotoxic protein from Ceratocystis fimbriata f. sp. platani. *J Biol Chem* . 1999;274(35):24959-24964. doi:10.1074/jbc.274.35.24959

53. Coil D, Jospin G, Darling AE. A5-miseq: An updated pipeline to assemble microbial genomes from Illumina MiSeq data. *Bioinformatics* . 2015;31:587-589. doi:10.1093/bioinformatics/btu661

54. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. *Trends Genet* . 2000;16(6):276-277. doi:10.1016/S0168-9525(00)02024-2

55. Sampedro J, Cosgrove DJ. The expansin superfamily. Protein family review. *Genome Biol* . 2005;6(12):242. doi:10.1186/gb-2005-6-12-242

56. Kende H, Bradford KJ, Brummell DA, et al. Nomenclature for members of the expansin superfamily of genes and proteins. *Plant Mol Biol* . 2004;55:311-314. doi:10.1007/s11103-004-0158-6

57. Krieger F, Möglich A, Kiefhaber T. Effect of proline and glycine residues on dynamics and barriers of loop formation in polypeptide chains. *J Am Chem Soc* . 2005;127:3346-3352. doi:10.1021/ja042798i

58. Serrano L, Neira JL, Sancho J, Fersht AR. Effect of alanine versus glycine in α-helices on protein stability. *Nature* . 1992;356:453-455. doi:10.1038/356453a0

59. Gräff M, Buchholz PCF, Stockinger P, Bommarius B, Bommarius AS, Pleiss J. The Short-chain Dehydrogenase/Reductase Engineering Database (SDRED): A classification and analysis system for a highly diverse enzyme family. *Proteins Struct Funct Bioinforma* . 2019;87(6):443-451. doi:10.1002/prot.25666

60. Vogel C, Pleiss J. The modular structure of ThDP-dependent enzymes.*Proteins-Structure Funct Bioinforma* . 2014;82(10):2523-2537. doi:10.1002/prot.24615

Tables

Table 1 : Numbers of homologous families (Hfams), protein entries, sequence entries and crystal structures in the different superfamilies (Sfam) of the ExED. Superfamilies 1 to 3 were named after the kingdom of their most abundant source organisms. Superfamily 4, named 'N-terminal domains', comprises only proteins containing the N-terminal expansin domain.

Sfam	**Sfam name**	**Hfams**	**Proteins**	**Seq**
1 2 3 4	Bacterial expansins Fungal expansins Plant expansins N-terminal domains	6 1 17 8	795 421 6,636 2,102	1,17
Total	Total	32	9,954	12,4

Table 2: The conserved amino acids or groups of amino acids according to the standard numbering scheme for the N- or C-terminal expansin domain, with the sequence of *Bacillus subtilis* (PDB accession 4fer) as reference sequence. All positions are listed separately for superfamilies 1 "Bacterial expansins", 2 "Fungal expansins", and 3 "Plant expansins" that are at least conserved to 70%. Positions marked in the standard numbering scheme as inaccurate are excluded (described in the Methods section). The last column names the function and the motif known from literature[9,55]. If a single amino acid is at least conserved to 70%, the conservation of the respective amino acid group is not mentioned. Amino acid groups: non-polar (A, C, F, G, I, L, M, P, V, W)[40]; polar (D, E, H, K, N, Q, R, S, T, Y)[40].

Standard position	Conserved amino acids in expansins from	Conserved amino acids in expansins from	Conserved amino acids in expansins from	Function and motif
	Bacteria	Fungi	*Viridiplantae*	
N-terminal expansin domain	N-terminal expansin domain	N-terminal expansin domain	N-terminal expansin domain	N-terminal expansin domain

Standard position	Conserved amino acids in expansins from	Conserved amino acids in expansins from	Conserved amino acids in expansins from	Function and motif
12 14 21 23 36 52 53 55 60 61.8 71 73 74 75 82 83 88 97	T (73%) Polar (76%) G (91%) Non-polar (92%) A (94%), G (4%) Non-polar (100%) G (100%) Y (70%), C (13%) G (92%), N (3%) D (97%), N (3%) P (93%), G (2%) E (73%), G (11%) D (99%) L (96%), M (3%) F (98%), Y (1%) G (96%)	T (71%), S (3%) Y (86%), T (1%) G (98%) C (99%) A (79%), C (19%) C (100%) G (100%) C (100%) G (70%), Y (21%) D (78%), N (21%) C (79%), T (19%) P (73%), G (19%) D (99%), N (3%) L (81%), M (9%) F (82%), W (16%) G (95%)	T (82%), A (4%) Y (85%), F (2%) G (96%) C (97%) A (83%), G (9%) C (99%) G (99%) C (98%) C (99%) C (96%), Y (1%) Polar (99%) C (79%), N (9%) P (76%), Y (5%) G (78%), N (6%) D (80%), V (8%) L (70%), M (26%) F (82%), W (9%) G (93%), S (1%)	T(F/W)YG-motif T(F/W)YG-motif GGACGYG-motif Disulfide bridge[14], GGACGYG-motif Disulfide bridge[14] Disulfide bridge[14] Disulfide bridge[14] Important for wall extension activity[10] Important for wall extension activity[10] Moderate role for wall extension activity[10] H(F/L)DL-motif, crucial for wall extension[10] H(F/L)DL-motif
C-terminal expansin domain 119 125 126 149 157 179	C-terminal expansin domain K (97%), Q (1%) W (63%), Y (30%) W (98%), Y (1%) W (88%), F (8%) Y (90%), W (3%) G (91%), H (5%)	C-terminal expansin domain K (89%), H (6%) Y (58%), N (15%) W (95%), F (3%) W (99%), Y (1%) Y (93%), P (4%) G (95%), K (3%)	C-terminal expansin domain Polar (94%) Y (61%), N (14%) F (57%), W (13%) W (93%), C (3%) G (99%), S (1%) G (76%), R (12%)	C-terminal expansin domain Important for wall extension activity[10] Binding to cellulose material[10] Binding to cellulose material[10] Binding to cellulose material[10]

Table 3: Known motifs from expansins in literature[9,55] and the newly suggested motifs for the N- and C-terminal expansin domains based on the conservation analysis performed in this study (**Tables S4** , **S5** , and https://doi.org/10.18419/darus-735). Sequence motifs known from literature are marked with a star (*). Polar residues are abbreviated with "p" and non-polar residues with "n". The expansins from *Viridiplantae* are separated into EXPA, EXPB, EXLA, and EXLB.

Standard positions	Bacterial expansins	Fungal expansins	EXPA	EXPB	EXLA	EXLB
Motifs in the N-terminal expansin domain 12 to 14 and 14.1 20 to 26 34 to 38 58 to 61 80 to 83	Motifs in the N-terminal expansin domain VpGP HLD*L	Motifs in the N-terminal expansin domain T(F/W)YG* GTAnS VpGn HLD*L	Motifs in the N-terminal expansin domain T(F/W)YG* GGACG*YG HFD* (L/I/V)	Motifs in the N-terminal expansin domain T(F/W)YG* GGACG HFD*L	Motifs in the N-terminal expansin domain GACG*YG	Motifs in the N-terminal expansin domain GACG(Y/F)G
Motifs in the C-terminal expansin domain 119 to 123 126 to 128 130 to 133 (and 134) 141 to 146 (including 143.1 and 143.2) 156 to 158 164 to 166 170 to 175	Motifs in the C-terminal expansin domain KpG(S/T)S QVRNH	Motifs in the C-terminal expansin domain KpG(S/T)S QVnN LEV-STDGD GGG VDVRVT	Motifs in the C-terminal expansin domain WGp LSFpVT	Motifs in the C-terminal expansin domain pGS WGA	Motifs in the C-terminal expansin domain YLA pGA	Motifs in the C-terminal expansin domain YLA (Y/F)GA

Figure legends

Figure 1 Functionally relevant positions in the expansin domains from the representative protein structure of *Bacillus subtilis* expansin *Bs* EXLX1 (PDB entry 4fer, chain B) are labelled with standard position numbers (numbering according to[13]) and shown as sticks. The substrate cellohexaose is depicted above the C-terminal expansin domain in green.

Figure 2 Protein sequence network showing the sequence space of the expansin sequences in the ExED belonging to the superfamilies 'Bacterial expansins', 'Fungal expansins', and 'Plant expansins'. All protein sequences presented in this network have a sequence length between 210 and 300 amino acids (**Figure S4**). The threshold for the nodes is 90% sequence identity (clustered with USEARCH) and the threshold for the edges is 50% pairwise sequence identity (determined by Needleman-Wunsch alignments). This network consists of 3504 nodes and 1,036,745 edges. With respect to the taxonomic lineages, the nodes from*Bacteria* , Fungi, *Viridiplantae* , and other origin are colored in red, orange, green, and white, respectively. The protein sequences of the fifteen biggest clusters belong to the following homologous families (Hfams) and expansin classifications: A (Hfams 9-20; expansin classification EXPA), B (21-22; EXPB), C (24-25; EXLB), D (23; EXLA), E (7, Fungi), F (3, 4; EXLX), G (1, 2, 4, 6; EXLX), H (3; EXLX), and I (7, Fungi). The bacterial sequences (red) in clusters A and D belong to *Streptomyces acidiscabies* (NCBI accession WP 050370046.1), *Kutzneria* sp. 744 (NCBI accession EWM10128.1, both Hfam 5, cluster A), and *Soehngenia saccharolyta* (NCBI accession TJX44964.1, cluster D).

Figure 3 Bivariate histogram of co-occurring HMMER bit scores of the N- and C-terminal expansin domains. The greyscale bar represents the relative frequency of the bit scores. The black diagonal line is the bisecting line.

Figure 4 Protein sequence network showing the protein sequence space of GH45 sequences from Pfam[45]

(accession PF02015) and the sequence regions annotated as N-terminal expansin domains from the superfamilies 'Bacterial expansins', 'Fungal expansins', 'Plant expansins', and 'N-terminal domains'. The colors representing the origin of the expansin sequences correspond to the scheme in **Figure 2** with GH45 sequences colored in blue. The threshold for the nodes is 90% sequence identity (clustered with USEARCH) and the threshold for the edges is 30% pairwise sequence identity (determined by Needleman-Wunsch alignments). This network consists of 4,031 nodes and 2,182,810 edges.

Figure 5 Protein sequence network showing the protein sequence space of CBM63 sequences from CAZy and the protein sequences of the superfamilies 'Bacterial expansins', 'Fungal expansins', and 'Plant expansins' with a sequence length between 210 and 300 amino acids (**Figure S4**). In contrast to the four big clusters from 'Plant expansins' (EXPA (A), EXPB (B), EXLA (C), and EXLB (D)), where no CBM63 sequences can be found, the clusters from the superfamilies 'Bacterial expansins' and 'Fungal expansins' show many connections to sequences of CBM63. The colors representing the origin of the expansin sequences correspond to the scheme in **Figure 2** with CBM63 sequences colored in cyan. The threshold for the nodes is 90% sequence identity (clustered with USEARCH) and the threshold for the edges is 50% pairwise sequence identity (determined by Needleman-Wunsch alignments). This network consists of 3,344 nodes and 844,280 edges.

Figures

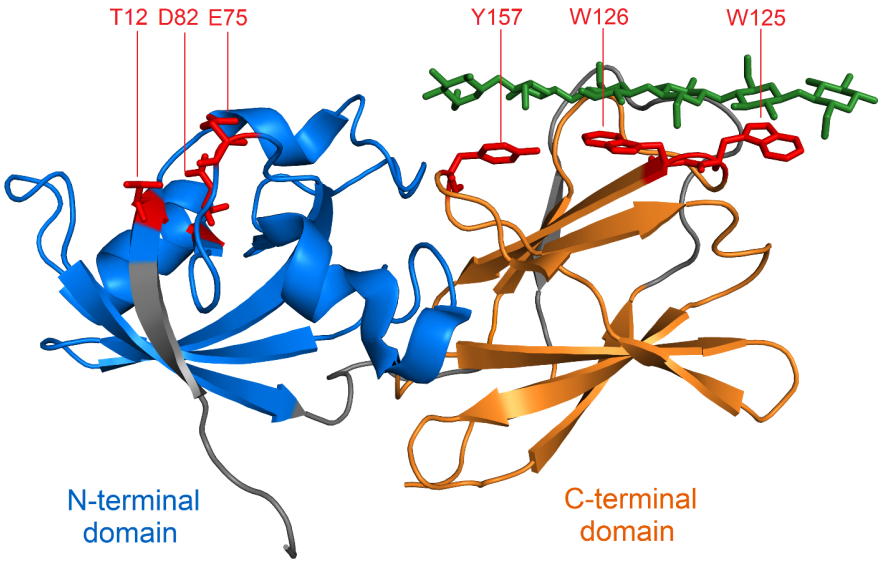

Figure 1

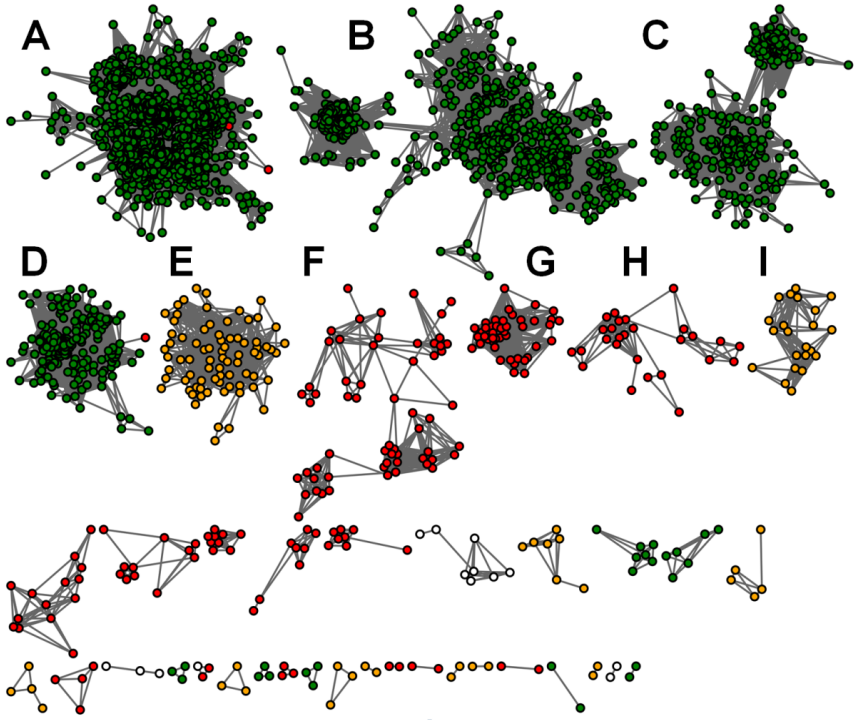

Figure 2

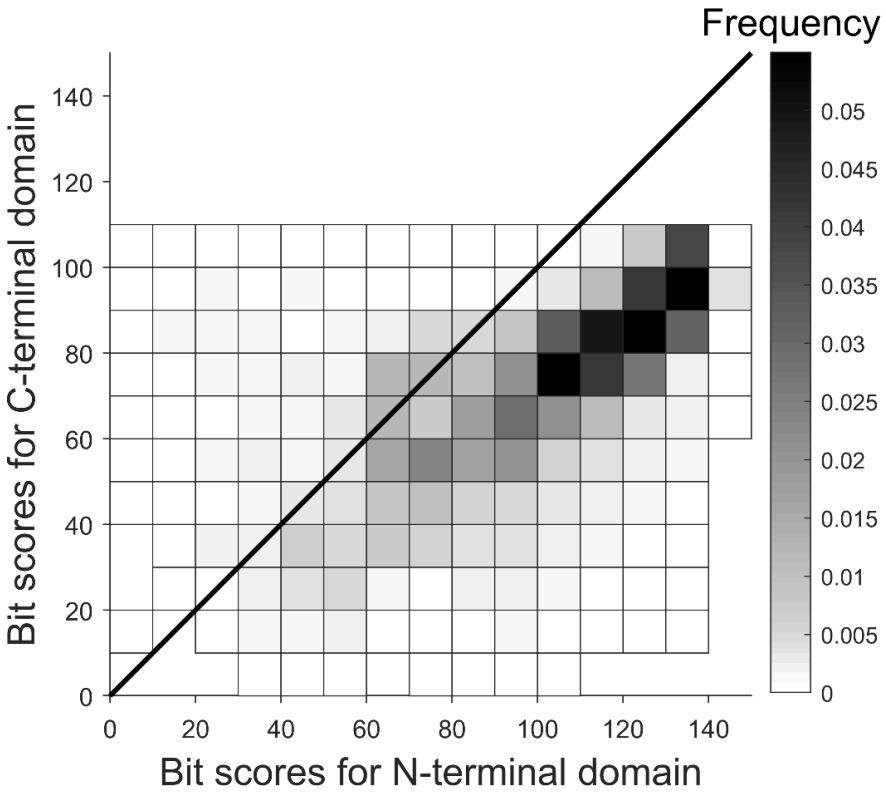

Figure 3

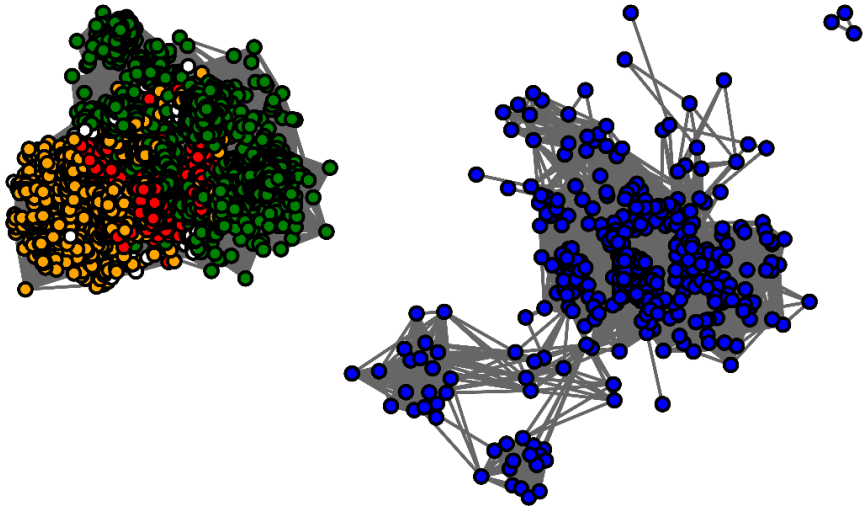

Figure 4

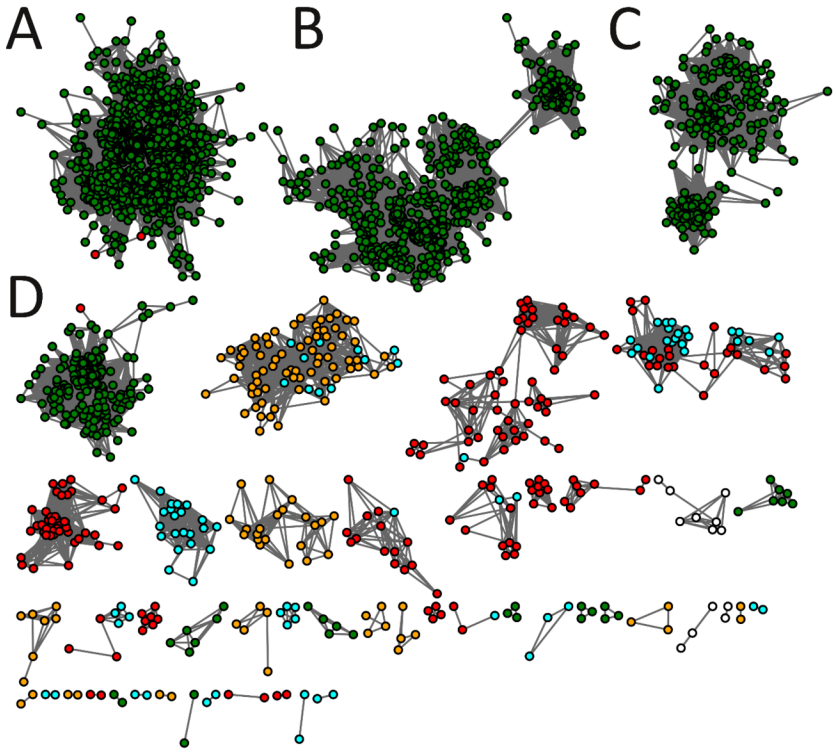

Figure 5