SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of Biotechnology and Biosciences

PhD program in Biology and Biotechnology

Cycle XXXII

Biochemical and biophysical analysis of two Antarctic lysozyme endolysins and *in silico* exploration of glycoside hydrolase 19 sequence space

Surname: Orlando Name: Marco

Registration number: 823497

Tutor: Prof. Marina Lotti

Coordinator: Prof. Paola Branduardi

ACADEMIC YEAR 2018/2019

Index

Abstract

Biodiversity of organisms and their genomic content is a valuable source of enzymes, some of which can be isolated and turned into biocatalysts, useful for more sustainable and efficient industrial processes.

Organisms thriving in constantly cold environments produce enzymes that may be more efficient in the cold and more thermolabile than enzymes from other organisms, and that display interesting features for the catalysis of several processes that require or are better at low temperature.

In the first part of this thesis, two glycoside hydrolases of family 19 (GH19), named LYS177 and LYS188, were identified in the genome of an Antarctic *Pseudomonas* strain and characterized. Even though most of the characterized GH19 are chitinases, LYS177 and LYS188 showed no chitinolytic activity, but were active as lysozymes with an optimum temperature of 25-35°C, and retained 40% of their highest activity at 5°C. The temperatures of midpoint unfolding transition were estimated to be 20°C higher than their optimum of activity. Based on these features and sequence analysis, LYS177 and LYS188 can be considered cold-active phage endolysins integrated in prophagic regions of the bacterial host. Moreover, the best performing of the two, LYS177, was active and structurally stable over several days only at 4°C, indicating it as a candidate for potential application on the preservation of food and beverages during cold storage.

In protein families, enzymes can rapidly acquire new specializations. Therefore, best practices should be implemented to select optimal candidates with the activity of interest and new, potentially promising, features.

Characterized GH19 enzymes showed an enhanced *in vivo* crop defence against chitin containing pathogens and antimicrobial potentialities.

In the second part of this thesis, the sequence space of the GH19 family was explored and a database was created to highlight non-described sequences potentially endowed with interesting variants.

Based on global pairwise sequence identity of all proteins available in public databases, GH19s were assigned to two subfamilies, the chitinases and the endolysins. Subfamilies were further split into homologous families, which differ in

the n° of characterized enzymes they harbour, in the taxonomical distribution, in the presence of accessory domains and loop insertions.

Despite this heterogeneity, a core consisting of 27 amino acids around the active site, including important substrate binding residues, was inferred to be conserved between GH19 subfamilies. Thus, this shared core is suggested to be associated to the GH19 capacity to bind sugars containing N-acetyl-glucosamine.

Moreover, specifically conserved positions in each subfamily alignment were identified to be a “signature” useful for predicting the substrate specialization of chitinases and endolysins, and to indicate possible outliers with different features.

The GH19 evolution was also investigated through molecular phylogeny to explain the observed sequence and structural plasticity: despite endolysins were divided in an higher number of homologous families, they remained in phages and their bacterial hosts, contrary to chitinases, which spread to both prokaryotic and eukaryotic taxa, and acquired at least four loop insertions; moreover, the GH19 chitinase catalytic domain passed from plants to bacteria by horizontal gene transfer in at least two cases.

In conclusion, the second part of this thesis shows how bioinformatic tools can be used to analyse the sequence space of a glycoside hydrolase family and extract information to help both experts and non-experts to optimize the discovery of new biocatalysts potentially applied in the field of human health and nutrition.

Riassunto

La biodiversità degli organismi e dei loro rispettivi genomi rappresenta una valida fonte di enzimi, alcuni dei quali possono essere isolati e convertiti a biocatalizzatori, utili in processi industriali più efficienti e sostenibili.

Gli organismi che vivono in ambienti costantemente freddi producono enzimi che possono essere più efficienti con il freddo e al contempo più termolabili di quelli provenienti da altri organismi, e che mostrano caratteristiche interessanti per catalizzare parecchi processi che richiedono o funzionano meglio a basse temperature.

Nella prima parte di questa tesi, due glicosil idrolasi della famiglia 19 (GH19), chiamati LYS177 e LYS188, sono stati identificate nel genoma di un ceppo antartico di Pseudomonas e caratterizzati.

Anche se molti enzimi GH19 sono chitinasi, LYS177 e LYS188 non hanno mostrato attività chitinolitica, ma sono risultati attivi come lisozimi con un optimum di temperatura a circa 25-35°C, e hanno mantenuto a 5°C il 40% della loro massima attività. Le loro temperature di "midpoint" di denaturazione termica sono state stimate essere 20°C più alte dell'optimum di attività. Sulla base di queste caratteristiche e su un'analisi di sequenza, LYS177 e LYS188 possono essere considerate endolisine attive a bassa temperatura, integrate in regioni profagiche del batterio ospite. Inoltre, l'enzima con le prestazioni catalitiche migliori, LYS177, è risultato attivo e stabile per parecchi giorni solo a 4°C, indicando la sua potenziale applicazione nella preservazione di cibi e bevande durante lo stoccaggio in frigo.

Nelle famiglie proteiche, gli enzimi possono rapidamente acquisire nuove specializzazioni. Delle "ottime pratiche" dovrebbero pertanto essere sviluppate per la selezione di candidati ottimali con le attività di interesse e nuove caratteristiche potenzialmente promettenti.

Gli enzimi GH19 caratterizzati hanno mostrato di essere in grado di migliorare *in vivo* la capacità difensiva delle colture contro patogeni contenenti chitina e di avere un potenziale antimicrobico.

Nella seconda parte di questa tesi, lo spazio di sequenza della famiglia GH19 è stato esplorato ed è stato creato un database per dare rilevanza a sequenze non ancora studiate e che possiedano delle varianti potenzialmente interessanti.

Sulla base di identità globale tra coppie di sequenze di tutte le proteine disponibili nei database pubblici, le GH19 sono state assegnate a due sottofamiglie, le chitinasi e le endolisine. Queste sottofamiglie sono state a loro volta divise in "homologous family", le quali differiscono nel n° di enzimi caratterizzati che contengono, nella tassonomia degli organismi, nella presenza di domini accessori, e in inserzioni chiamate "loop".

Nonostante questa eterogeneità, è stato identificato un "core" di 27 amminoacidi attorno al sito attivo, conservato tra le sottofamiglie di GH19, e comprendente residui importanti per il legame al substrato. Pertanto, si può suggerire che questo "core" condiviso sia associato alla capacità delle GH19 di legare zuccheri contenenti N-acetyl-glucosammina.

Inoltre, delle posizioni conservate specificamente nell'allineamento di ogni sottofamiglia sono state identificate per essere usate come "signature" nella predizione della specificità di substrato in chitinasi ed endolisine, e per indicare possibili "outlier" con caratteristiche diverse.

Anche l'evoluzione delle GH19 è stata studiata mediante filogenesi molecolare per spiegare la plasticità di sequenza e di struttura: nonostante le endolisine siano state divise in più "homologous family", sono rimaste nei fagi e nei loro ospiti batterici, contrariamente alle chitinasi che si sono diffuse sia in *taxa* procariotici che eucariotici, e hanno acquisito almeno quattro inserzioni di "loop"; inoltre, il dominio catalitico delle chitinasi GH19 è passato dalle piante ai batteri mediante trasferimento genico orizzontale in almeno due casi.

In conclusione, la seconda parte di questa tesi mostra come i "tool" bioinformatici possano essere usati per analizzare lo spazio di sequenza di una famiglia di glicosil idrolasi al fine di estrarre informazioni utili sia ad esperti che a non-esperti per ottimizzare la scoperta di nuovi biocatalizzatori potenzialmente applicabili nel campo della salute e dell'alimentazione umana.

Abbreviations

NGS: next generation sequencing; **ORF**: open reading frame; **CAZy**: carbohydrate active enzyme database; **CAZymes**: carbohydrate active enzymes; **GH**: glycoside hydrolase; **GH19**: glycoside hydrolase family 19; **GlcNAc**: N-acetyl-glucosamine; **HGT**: horizontal gene transfer; **COS**: chitooligosaccharide; **CBM**: carbohydrate binding module; **CAE**: cold-active enzyme; $\mathbf{T_{opt}}$: optimum temperature; $\mathbf{T_m}$: midpoint denaturation temperature; **PBM**: peptidoglycan binding module; **GH19ED**: GH19 engineering database; **ELYS**: endolysin; **CHIT**: chitinase; **CLP**: chitinase-like protein; **OM**: outer-membrane; **AN**: accession number; $\mathbf{pH_{opt}}$: pH optimum; **HEWL**: hen egg white lysozyme; **pNP-chitobioside**: 4-nitrophenyl N,N′-diacetyl-β-D-chitobioside; **CA**: Chitin Azure; **CD**: circular dichroism; **PDB**: protein data bank; **CHIT**: chitinase; **ELYS**: endolysin; **HMM**: hidden Markov model

1.Introduction

1.1 Enzyme discovery

1.1.1 Enzyme discovery: from the past to the era of next generation sequencing

Biocatalysts are biological enzymes applied in the catalysis of biochemical reactions, applied in many different biotechnological processes for research or industrial purposes.

Enzymes have been viewed for more than 20 years as the 'third wave' of biotechnology [1]: they have the advantage to be efficient and selective in the chemistries they accelerate, catalyse the production of relatively pure products, thus minimizing waste, and they can carry out regio-, chemo- and stereospecific reactions that are challenging for conventional chemistries. Reactions catalysed by enzymes are carried out in conditions usually compatible with biological processes, and with much less toxic reagents [2]. In this way biocatalysts support high levels of safety, low energy consumption, and overall environmentally friendly production procedures [3].

Up to date, a small number of industrial biotechnological processes exploiting biological catalysts have been established, mainly in detergency and food processing industries. A limiting step in increasing the available biocatalytic processes is the search and identification of enzymes that fit the reaction of interest. Indeed, one of the main cause is that only a minor fraction of microbial diversity ($\approx$ 1%) can be cultured and isolated [4, 5] to directly assay the enzymatic activities of their crude cellular extracts.

A valuable alternative is represented by high throughput assay systems applied to purified recombinant enzymes expressed in microbial hosts optimized for laboratory conditions [6]. The purification step to obtain a sufficient amount of the soluble and stable form of the enzyme of interest is an issue in some case [7], hampering its use at industrial scale.

Since some enzymes do not tolerate the conditions required for the industrial process, and it could be that no enzyme is known for the desired application or it is not efficient enough to justify its use from an economical point of view, not every biocatalyst is useful [7, 8].

At the end of the past century, high interest was triggered by the discovery of enzymes from organisms living in extreme environments, the so-called extremophiles [9], capable to grow under extremes of temperature, pH, pressure, or salt concentration. Enzymes from extremophiles, so called extremozymes, can potentially work under harsh conditions, solving the stability problem encountered by biocatalysts in industrial processes, or being used to re-design existing processes based on mesophilic enzymes [10, 11]. In paragraph **1.5** of this thesis a more comprehensive overview will be provided for enzymes from psychrophilic organisms.

The recent use of extremozymes relies on advances in the technologies for the cultivation of extremophiles and on the ability to express genes from extremophiles into conventional hosts systems to produce recombinant enzymes under milder, less expensive growth conditions [12].

In spite of the recent progress in culturing techniques, most extremophiles cannot be grown using traditional protocols, and the amount of DNA isolated from low biomass collected from environments hostile for life is not always sufficient for the amplification and isolation of clone libraries from which isolate the gene coding for the enzyme of interest [12]. Moreover, even when the microorganism could be isolated, not all enzyme activities in the medium or crude extracts can be detected or elicited by substrate induced gene expression [13]. Therefore, the identification and isolation of new organisms producing the biocatalysts of interest is still a time-consuming step and the screening of recombinant clone library collections may be not the best choice.

Even if this hurdle has been partly overcome by automatization of the processes of high throughput library screening [14], different approaches have been developed to ensure a sustained search for new biocatalysts [15], which exploit Next Generation Sequencing (NGS) methods, that ease and reduce exponentially time and costs of sequencing starting from smaller amounts of DNA sample [16], combined by the use of informatic tools for the prediction of their functions. In this way, it has been possible since more than a decade to look directly on the enzyme sequences isolated form the genome of an organism or even a collection of DNA fragments from an

environmental sample (metagenome). In the meanwhile, the increased computational power and the capacity to store a huge amount of biological sequences in public databases enhanced the speed of protein identification and analysis, through fast *in silico* predictions of enzymes to select as candidates for further characterization [17]. In subsequent steps, if these candidates fit the activity of interest, different variants of the enzyme and the substrate can be tested in different conditions, up to the development of a new biocatalyst useful in an industrial process (**Fig. 1.1**).

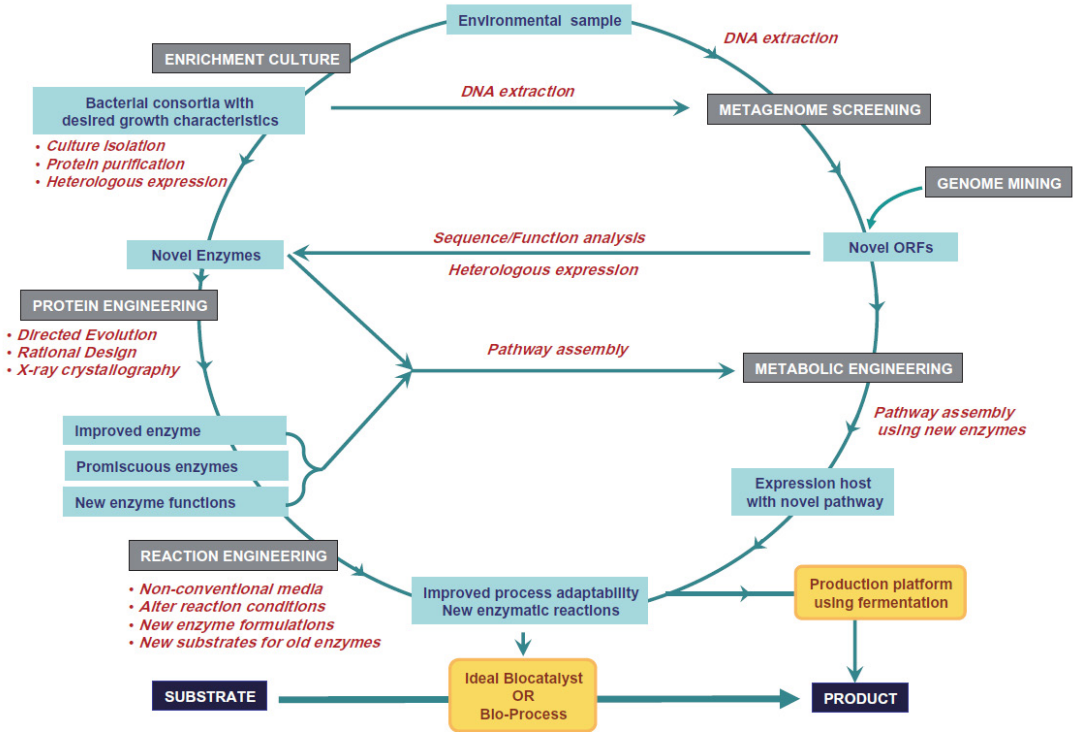


Fig. 1.1 Schematic representation of a biocatalyst discovery and implementation cycle starting from environmental DNA. The metagenome or an isolated genome are mined to search for novel open reading frames to analyse, screen for the desired functions, and eventually improve the characterized enzyme up to apply it in a bio-process for the industrial biotechnology sector. Reproduced from [15].

Such advancements have lead to previous time-consuming testing of large libraries codifying for unknown proteins to be a superseded approach [18], and have paved the way for the development of bioinformatic tools to obtain accurate predictions of protein functions from coding genes (a process named *genome mining*). Numerous studies have benefit from this approach applied to metagenomes, yielding enzymes with potential for biocatalytic applications, such as lipase [19, 20], oxidoreductase [21], amidase [22], amylase [23], nitrilase [24], β-glucosidase [25, 26], decarboxylase [27], epoxide hydrolase [28], cellulase [29], and others. Moreover, considering the drawbacks to retrieve enzymes from extreme environments through classical culturing procedures, *genome mining* has emerged as an opportunity for pushing the discovery of novel extremozymes [30].

1.1.2 Genome mining: automatic annotation pitfalls

NGS methods have increased the amount of genome assemblies, supporting the recent development of standardized *in silico* protocols for the automatic annotations of eukaryotic and prokaryotic genomes [31-36].

While slightly different approaches can be used (an example of pipeline is shown in **Fig. 1.2** [34]), all methods assign functions to open reading frames (ORFs, coding genes identified in the target genome by different heuristics) based on the same concept: in a first step, algorithms such as BLAST and its variants [37], and most recently PSI-BLAST [38] and HMMer [39] (developed to increase the search sensitivity) are used to find significantly similar matches with sequences contained in a reference public database (i.e., NCBI, https://www.ncbi.nlm.nih.gov/; Uniprot, https://www.uniprot.org/, etc…); in a second step, a list is generated (from the best to suboptimal matches) on the basis of an estimated statistical significance with respect to random matches. Then, the reference sequence with the most significant similarity is used to transfer the annotation to each ORF.

In the case a protein is modular and can be subdivided in more than one protein domain, the comparison of its ORF with the sequences present in a database could be

not straightforward to predict its function as a whole, because each domain (yet or not characterized) can exhibit a specific function that might be unrelated to other domains in the same protein (i.e., a non-enzymatic domain that binds the substrate and a catalytic domain that catalyses its hydrolysis, separated by a flexible linker region).

The combinations of domain architectures in proteins can be highly plastic, reflecting the great diversity of modular forms that exist in nature. This is the result of domain duplication and recombination events during protein evolution [40], and several known examples suggest that such processes can lead to new combinations with new functions [41]. Therefore, ORFs coding for a multi-domain protein annotated by relying on a statistically significant local similarity with a known reference protein, might result in a wrong or at least incomplete function assignment. Some recent tools and databases have been developed to tackle some of these problems by allowing the analysis of the domain composition of proteins, instead of directly comparing protein sequences. Examples of these domain databases are SMART [42], PFAM [43], InterPro [44] and Gene3D [45].

Anyway, the rapidly increasing number of sequences in databases required a rethinking the utility of the definition of homologous relationships among protein domains: statistically significant local or global sequence similarity among two proteins may be due to a more or less recent common ancestor, but does not mean they carry out the same molecular function [46], regardless if their domains are orthologues (deriving from modifications accumulated in different species since the speciation event) or paralogues (deriving from two or more different events of duplications in the genome of the same species).

Substrate promiscuity [47] and the existence of multiple substrate specializations in the same protein are the major mechanisms responsible for the observed mismatch between statistical similarity in primary sequence and catalysed reaction in homology-based predictions of enzymatic functions. Enzymes can also evolve toward the same activity from completely different ancestral folds or by fusion of domains from different families. On this ground, they could be no more able to perform their

original activity, which is predicted on the basis of the most similar sequences in the database [48]. The evolutionary bases of these processes have been studied intensively in the last 10 years [49-52].

Moreover, as an annotation is transferred by similarity to a protein usually not characterized and whose annotation might also be not correct, there is a certain probability that chains of mis-annotations can be propagated inside a database [53], considering that generally it is not possible to visualize how function predictions were assigned up to the most similar characterized protein [54].

Even if sequence-based functional predictions are inaccurate, and in some cases even wrong, they still provide valuable guidelines for experimental studies and remain the best approach to start the functional annotation of uncharacterized proteins from newly sequenced genomes [51].

In order to overcome the limitations presented in this paragraph, new approaches should be developed and the predicted putative function/s of the enzymes of interest must be evaluated in wet lab by effective screening systems represented by qualitative or quantitative enzyme activity assays [55].

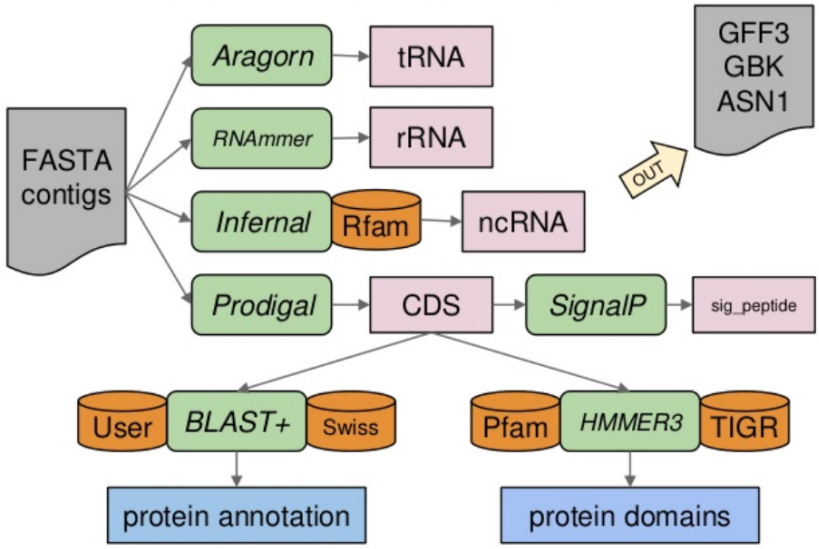


Fig. 1.2 Simplified representation of the prokaryotic genome annotation pipeline performed by the software package Prokka [34]. "CDS" are the open reading frames identified from "FASTA contigs", the collection of genome/metagenome sequences. Taken from a presentation published at https://www.slideshare.net/torstenseemann/prokka-rapid-bacterial-genome-annotation-abphm-2013.

1.1.3 Evolution-aware approaches to handle activity predictions

As explained in the previous paragraph, evolution played an important role in decoupling sequence similarity and protein functions. The idea behind considering evolution in predicting protein functions arose at the end of the last century, with the creation of the clusters of orthologous groups of proteins database (COG, [56]). However, at that time the computational power was insufficient for the use of model-based statistical inference on many sequences at once.

In the last years, the use of phylogenetic tools for reconstructing genealogies of protein families at domain level has been used to correlate the most significant evolutionary trajectories with observed patterns of functional diversity. Major mechanisms of protein evolution include events of duplication, horizontal gene transfer, recombination, genetic drift of redundant copies [40, 57], and co-evolutionary epistatic effects (within the same of among different proteins), which created constraints in possible evolving trajectories and opportunities for new functions [58]. Taking these events into account, it is possible to predict if candidate enzymes, close or far from known references, are endowed with potentially interesting properties to be tested by using different types of substrates.

For example, model-based genealogical reconstructions of a collection of homologous proteins can be used to estimate the relative evolutionary rate of different regions in the protein sequence, and to highlight sites that have undergone the same selective pressure [59, 60]. Considering also the protein structure and other features measured by experiments or simulations, correlated sites can be recognized important for substrate specificity. In this way, they can be used as signatures to assign specific functional properties.

Moreover, several conceptually different tools have been implemented for optimizing protein function prediction based on the phylogenomic context, in particular for prokaryotic genomes, as reviewed in [61]. Among these, some methods make use of phylogenetic-aware distribution patterns (**Fig. 1.3** Co-occurence): functionally related genes tend to co-evolve or to be lost in concert in the genome of different organisms. On the contrary, if a subset of genes co-occur in two different

genomes without co-evolving, the functional interactions between the corresponding gene products tend to be less dominated by physical interactions [62]. Alternatively, a gene that is never found when another one completely different is present, it could mean the protein products of the two genes carry out the same function even if they do not possess any sequence similarity (**Fig 1.3** Anti-correlation): in this way, if the function of one of the two is known, it can be used for predicting the function of the other.

One should be aware that in some protein families and superfamilies the same catalytic mechanism can be reused to easily acquire new substrate specificities [63].

These situations are difficult to be predicted, even if looking to protein evolution. Therefore, the protein structure should be integrated to improve the phylogenetic reconstructions [64, 65], and to calculate the affinity of the active site for different docked substrates in the three dimensional dynamic network of interactions [66, 67].

Some well documented cases [68-74] demonstrated how exploiting sequence-based phylogenetics and evolutive patterns, in integration with structural und functional information, allowed to improve the reliability of enzymatic activity predictions over the results obtained by sequence-similarity based methods.

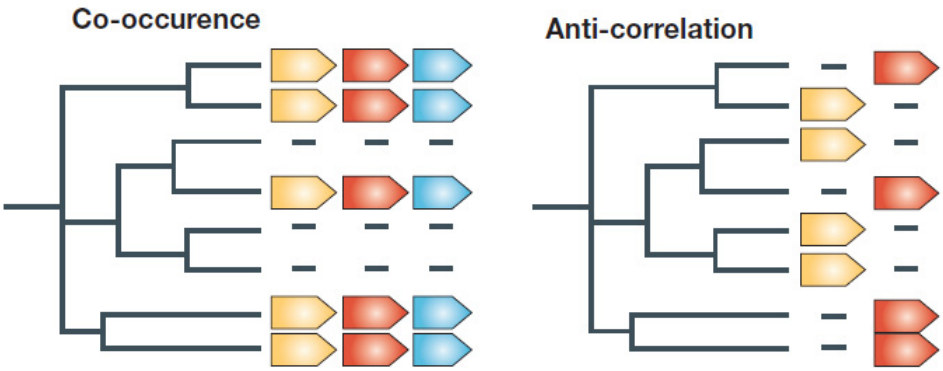


Fig. 1.3 A schematic representation of a genomic context dependent method for the prediction of coding gene functions: the phylogenetic reconstruction is used to map co-occurrence of functionally associated homologous genes or anti-correlation of non-homologous genes that most probably are functionally equivalent. Modified from [61].

1.2 Carbohydrate active enzyme database (CAZy) and CAZymes discovery

1.2.1 Introduction to CAZy

The enzymes acting on glycoconjugates, oligo- and polysaccharides, designated as carbohydrate active enzymes (CAZymes), probably constitute one of the most structurally and functionally diverse set of proteins. As carbohydrate diversity [75] exceeds by far the number of protein folds, CAZymes have evolved from a limited number of progenitors by acquiring novel specificities at substrate and product level, representing an example of how evolutionary processes challenge the functional predictions based on sequence similarity methods.

The Carbohydrate-Active Enzyme (CAZy) database (http://www.cazy.org/) is a knowledge-based resource available since 1998, specialized in the classification of CAZymes into families based on similarity in their primary sequence; each family is then associated to one or more catalysed reaction, to a catalytic mechanism and to a protein fold [76].

In CAZy there are 5 major classes based on the general type of activity: glycoside hydrolases (GHs), which catalyse the hydrolysis and/or transglycosylation of glycosidic bonds; glycosyl transferases, which catalyse the synthesis of glycosidic bonds from phospho-activated sugar donors; polysaccharide lyases, which cleave the glycosidic bonds of uronic acid-containing polysaccharides by a β-elimination mechanism; carbohydrate esterases, which remove ester based modifications in mono-, oligo- and polysaccharides and thereby facilitate the action of GHs; carbohydrate-binding modules, which are non-enzymatic domains known to potentiate the activity of many enzymes from the classes described above by binding to the substrate and increasing its prolonged interaction with the catalytic domain.

Each protein sequence in CAZy includes the annotations in publicly available sources (NCBI), its family classification system [77], and known/predicted structural and functional information. A feature of CAZy is that new families are built around new seed sequences biochemically characterized, based on references extracted automatically from individual accessions with ProFal [78], or entered manually.

In the last years, the increasing number of sequences led to the reorganization of CAZy interface design and contents [79]: the aesthetics changed, some GH families were included into clan sharing the same fold, some families were further divided into subfamilies to improve the prediction of substrate specificity [80-82]. A new CAZy class named 'Auxiliary Activities' was also created to accommodate mainly lytic polysaccharide mono-oxygenases responsible for redox enzymatic conversion of lignocellulosic material in concert with other CAZymes [83].

1.2.2 CAZymes and rational exploration of protein sequence space

Large-scale phylogenetic comparisons of microbial sequenced genomes highlighted both the relatedness of CAZymes involved in polysaccharide degradation and the variability of the domain organization in GHs from closely related organisms [84, 85]. The biochemical characterization of these similar CAZymes present in many copies in a genome showed that subtle variations exist in substrate specificity, enzymology, and regulation [86]. Moreover, the skewed taxonomic distribution of GH domains permitted to correlate specific polysaccharide degrading capacities to specific lineages of microorganisms, inhabiting certain environments [87, 88].

Despite these studies showed how effectively CAZymes can be described, the gap between the number of sequences and the number of biochemically or structurally characterized CAZymes (**Fig. 1.4**) was growing up to the present due to genomic data resulting from NGS, combined with the much lower pace of experimental and structural characterization [79].

In the same GH family, the often occurrence of enzymes that act on different substrates remain a significant problem for their automated functional annotation. This can sometimes be overcome by subfamilies defined on the basis of known substrate specificities, but for many families there is insufficient information to allow a complete unsupervised automated substrate prediction [79].

Moreover, CAZy database does not provide a tool for the direct addition and annotation of external sequences neither it is designed to ease the extraction and the analysis of information at large scale [89]. It is therefore necessary to introduce new

methods to explore the sequence and structure space, especially when dealing with multi-specific enzyme families from which only few sequences have been characterized and no subfamily classification is provided [90].

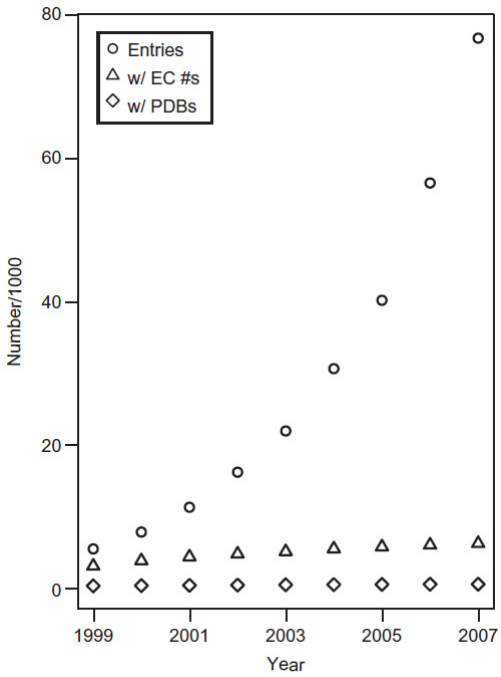


Fig. 1.4 The number of proteins containing CAZy modules (years 1999-2007) is represented with open circles, the number of enzymatically characterized proteins with triangles and that with solved structures with open diamonds. Reproduced from [76].

A possible solution could be looking to transcriptomics, proteomics, and metabolomics, that can reveal useful relationships between genes or proteins, but do not directly assign function or substrate specificity to hypothetical enzymes.

Another option is the development of systems for the easy visualization of CAZymes sequence space in each CAZy family/clan. In this way, it would be possible to immediately evaluate the distance in the sequence space between non-described sequences and the experimentally characterized enzymes [91]; at the same time, the "empty" regions of the sequence space, far from any known enzyme, would be interesting for the selection of candidates to experimentally screen. A similar strategy, based on the rational bioinformatic selection of CAZymes from subfamilies without characterized reference sequences and with low similarity (<20% identity) to known

families, was recently done [90], and the experimental screening showed that this selection procedure was effective.

The creation of sequence similarity networks [92] (**Fig. 1.5**) can also provide both a quantitative representation of sequence similarity relationships among sub-clusters in the sequence space, and an easy visualization of such relationships, that can be integrated with phylogenetic trees and known biochemical properties, permitting the analysis and visualization of much larger sets of sequences than trees [93, 94]. Information on other features (i.e., accessory domains, structural motifs, etc…) can also be visualized on these networks, providing a comprehensive overview of diversity and evolution within enzyme families and superfamilies.

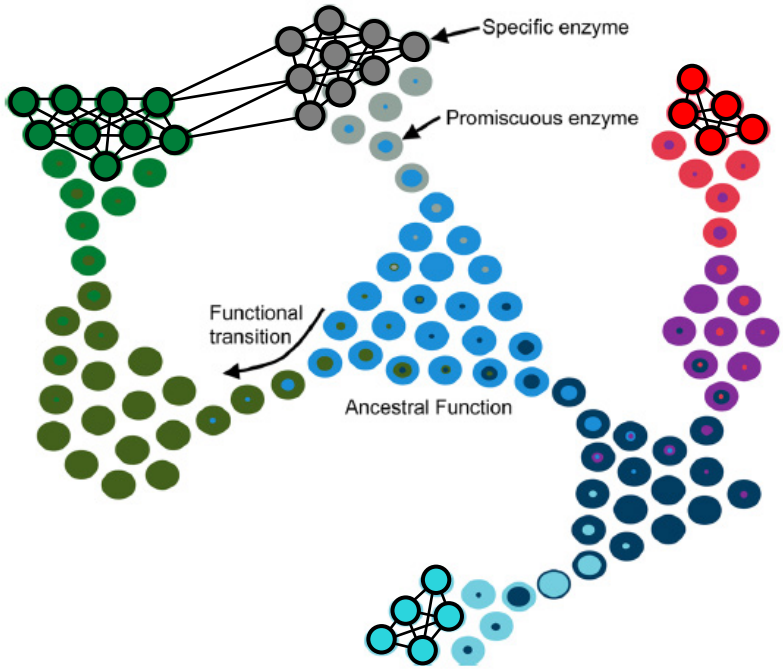


Fig. 1.5 Schematic representation of sequence similarity networks (each sequence is a node represented by a circle) plotted onto a scheme of the evolutionary process by which functional divergence occurs within a theoretical enzyme superfamily. Edge length are proportional to the distance in the sequence space, and absent under a certain threshold of pairwise sequence identity between enzymes (nodes). Different colours represent different functions. Inner circles mark promiscuous activities. The functional divergence from a common ancestor (light blue) occurs via the recruitment of promiscuous activities and the evolutionary optimization of these functions to generate new specialized enzymes. Modified from [94].

1.3 Glycoside hydrolases (GHs)

1.3.1 General hydrolytic mechanism

As a consequence of polysaccharide diversity, there is great variety among the GH families, which are around half of the total number of sequences in CAZy [76]. Despite this huge diversity, the hydrolysis of the glycosidic bond is usually catalysed by two amino acids, one acting as a general acid/base (proton donor) and the other as a nucleophile [95]. Depending on the spatial position of these catalytic residues, hydrolysis occurs via two catalytic mechanisms: retention or inversion. These names are due to the fate of the anomeric configuration of the carbon involved in the bond at the reducing saccharide end, after the hydrolysis.

In retaining enzymes, the nucleophilic catalytic base is close to the sugar anomeric carbon. This base, however, is more distant in inverting enzymes, which accommodate a water molecule for the nucleophilic attack between the base and the sugar. This difference results in an average distance between the two catalytic residues of 5.5 Å in retaining enzymes, as opposed to 10 Å in inverting [96]. In some cases, the catalytic nucleophile is not provided by the enzyme, and is replaced by the acetamido group at C-2 of the substrate [97]. A completely unrelated mechanism has been demonstrated recently for two families of GHs utilizing NAD+ as a cofactor [98, 99].

1.3.2 GH classification

There are over 160 GH families divided in 18 clans and 8 types of folds (http://www.cazy.org/Glycoside-Hydrolases.html). Despite this diversity in terms of sequences and structures, the overall topologies of the active sites fall into only three general classes (**Fig. 1.6**), regardless of whether the catalytic mechanism is inverting or retaining [95].

I. The pocket (**Fig. 1.6A**): optimal for the recognition of a saccharide non-reducing extremity and typical of β-galactosidase, β-glucosidase, sialidase and neuraminidase, and of exopolysaccharidases such as glucoamylase and β-amylase. Such

exopolysaccharidases are active on substrates with a large number of available free chain ends, but not very efficient on fibrous substrates such as native cellulose, which has almost no free chain ends.

II. The cleft (**Fig. 1.6B**): this 'open' structure allows the random binding of several sugar units in polymeric substrates and is commonly found in endo-acting polysaccharidases, such as lysozymes, endocellulases, chitinases, α-amylases, xylanases, β-1,3-1,4-glucanases and β-1,3-glucanases.

III. The tunnel (**Fig. 1.6C**): this topology arises from the previous one when the protein evolved long loops that cover part of the cleft. The resulting tunnel enables a polysaccharide chain to be threaded through it [100], permitting these enzymes to release the product while remaining firmly bound to the polysaccharide chain. Therefore, the conditions are created for processivity, which is probably a key factor for the enzymatic degradation of insoluble crystalline sugars. In either case it should be noted that, depending on the mechanism (inverting or retaining) and the exact position of the cleavage point, the directionality of the enzyme motion along the chain may change. For instance, cellobiohydrolase II of *Trichoderma reesei* proceeds towards the reducing end of cellulose chain, whereas the reverse was suggested for cellobiohydrolase I from the same organism [101].

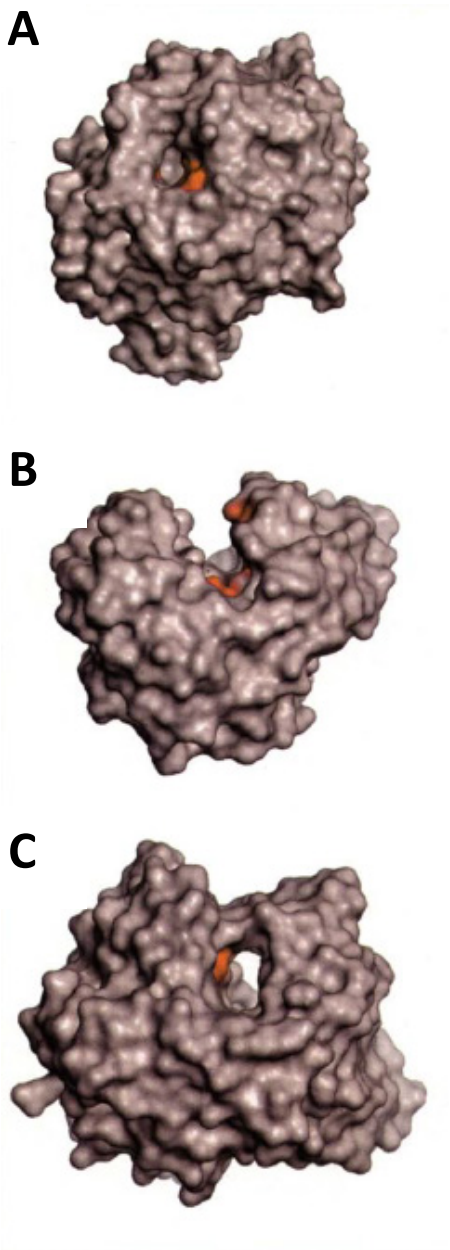


Fig. 1.6 The three types of active site found in GHs. **A**: the pocket (glucoamylase from *Aspergillus awamori,* PDB code: 1GLM). **B**: the cleft (endoglucanase E2 from *Thermobifida fusca,* PDB code: 2BOD). **C**: the tunnel (cellobiohydrolase II from *Trichoderma reesei,* PDB code: 3CBH). The proposed catalytic residues are shaded in red. Reproduced from [95].

1.4 Glycoside hydrolase family 19 (GH19)

1.4.1 Activity, specificities and catalytic mechanism

Chitinases (EC 3.2.1.14) and lysozymes (EC 3.2.1.17) break down chitin and peptidoglycan polymers, respectively, by hydrolyzing the glycosidic covalent bonds between their monomeric subunits. For this reason, they belong to GHs families [95].

Chitin is an insoluble homopolymer of β-(1–4)-linked N-acetylglucosamine (GlcNAc), the second most abundant sugar polymer in the biosphere (after cellulose) [102]. Chitinases can be found in any type of organism, with biological roles spanning from defense against chitin-containing Fungi, to the use of chitin as carbon source and for chitin recycling and morphogenesis in organisms in which this polysaccharide forms part of their cells or body (like cuticle for insects and cell wall for Fungi) [103, 104].

Peptidoglycan is a heteropolymer of β-(1–4)-linked GlcNAc and N-acetylmuramic acid, the sugar component found in the cell wall of Eubacteria [105]. The lysozyme families are considered to belong to a superfamily sharing the same structural core [106], and have been studied as antimicrobial specialized enzymes, in particular in animals [107].

GH19 is a family, defined in CAZy, that collects enzymes characterized both as chitinases and lysozyme endolysins. “Endolysin” is a generic term used to indicate many different enzymes produced by bacteriophages at the end of their replication cycle to degrade the peptidoglycan of the bacterial host from within, resulting in cell lysis and release of progeny virions [108]. Endolysins can be named differently with respect to their hydrolysis target (**Fig. 1.7**).

GH19s have a single displacement catalytic mechanism causing the inversion of the anomeric C-1 (**Fig. 1.8**) [109, 110]; their chitinolytic activity is most probably endo-type, because of the “cleft” active site shape [95] and the predominance of hydrogen bonds in stabilizing the interactions in enzymes complexed with chitooligosaccharides, COSs [111]; the endolysin activity is lysozyme-like [112, 113].

Although GH19 chitinases and endolysins are specialized in the hydrolysis of two substrates, CAZy does not provide any subfamily classification for them

(http://www.cazy.org/GH19.html) and, despite lysozyme activity is reported, only chitinases are mentioned in the notes. Nevertheless, the GH19 sequence motifs described in literature are valid only for the generic annotation of GH19 domain [114], and there is not any study comparing the conserved and specific features of GH19 chitinases and endolysins to date.

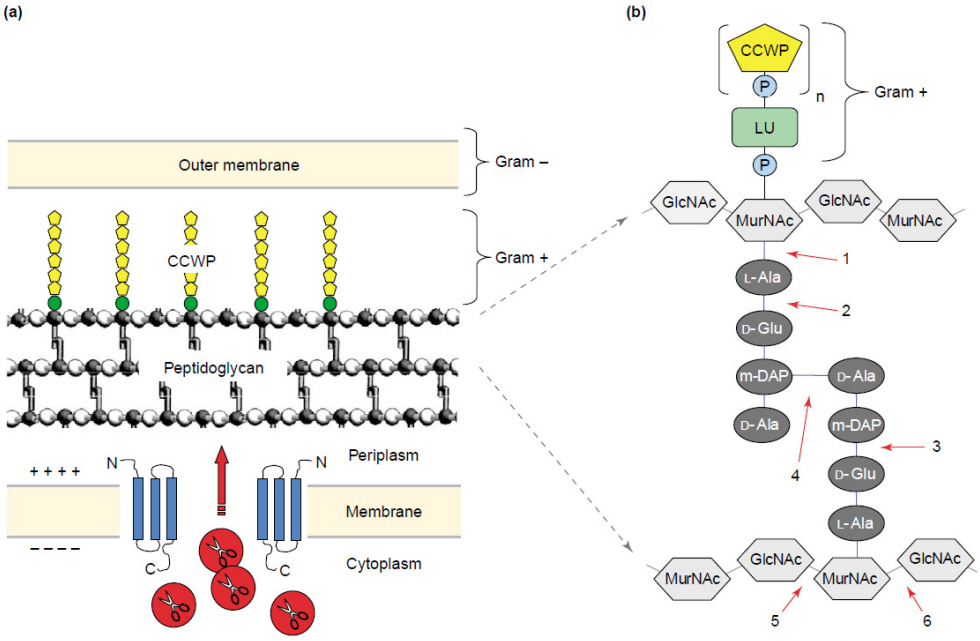


Fig. 1.7 Bacterial cell wall structure and endolysin targets. **A.** Schematic representation of the bacterial cell wall, and of how phage endolysins gain access to their substrate. Holin proteins (blue) insert themselves into the cytoplasmic membrane and can oligomerize, thereby forming membrane lesions. Endolysins (red) cross these pores to access the peptidoglycan. Lysis of Gram-positive cell walls is possible from the outside. In Gram-negative cells, the outer membrane is an efficient barrier to prevent lysis by free endolysins. **B.** The bonds potentially attacked by endolysins with different specificities are indicated by numbers: 1, N-acetylmuramoyl-L-alanine amidase; 2, L-alanoyl-D-glutamate endopeptidase; 3, D-glutamyl-m-DAP endopeptidase 4, interpeptide bridge-specific endopeptidases; 5, N-acetyl-β-D-glucosaminidase; and 6, N-acetyl-β-D-muramidase (also known as muramoylhydrolase and *lysozyme*), the type of activity performed by GH19.
CCWP, carbohydrate cell wall polymer; GlcNAc, N-acetyl glucosamine; LU, linkage unit; m-DAP, meso-diaminopimelic acid; MurNAc, N-acetyl muramic acid; P, phosphate group. Reproduced from [115].

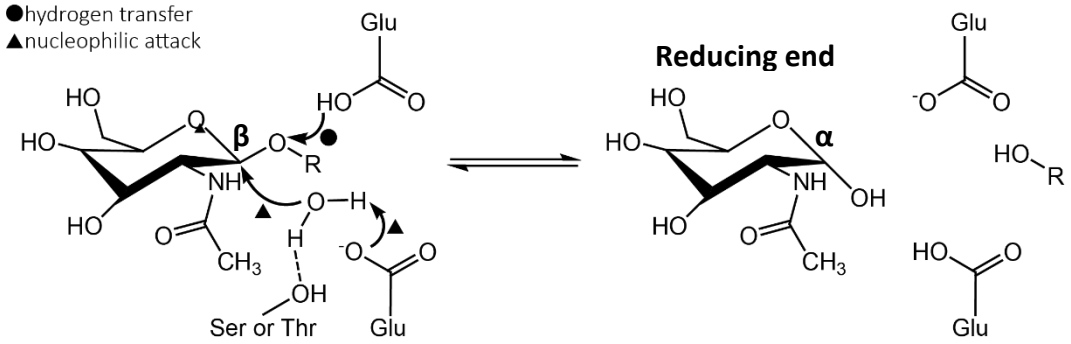


Fig. 1.8 The single displacement hydrolytic mechanism of GH19s [110]. One acidic, one basic glutamate and a serine (or threonine) for water placement are generally present in the active site. The reducing end of the hydrolysis product has inversion of the anomeric configuration from β to α.

1.4.2 GH19 classification

GH19s were first discovered in the late 1980s in plants, and classified as class I, II and IV chitinases [116-119].The observations made on the structures available since the beginning of 1990s revealed that these enzymes have a catalytic core spanning a deep catalytic cleft connecting two lobes rich in α-helices, while six surface loops around the cleft (1, 2, 3, 4, 5 and C-terminal [120]) can be present or absent depending on the enzyme (**Fig. 1.9A**).

Chitinases of classes I and IV carry an accessory N-terminal carbohydrate binding module 18 (CBM18). Class IV chitinases do not possess some of the loops, causing a reduced length of the catalytic cleft and a different substrate binding mode [111, 121, 122]. This explains why class IV chitinases have been recently called “loopless”, contrary to “loopful” term used for other GH19 [123]. Chitinases of class II, contrary to classes I and IV, do not possess any accessory CBM.

Class IV “loopless” chitinases were also found in Actinobacteria, for which horizontal gene transfer (HGT) of GH19 genes has been suggested from plants [124, 125]. Characterized bacterial GH19 chitinases possess an N-terminal CBM5/12, which is different from that found in GH19 plant chitinases [114, 125, 126].

These GH19 classes of chitinases were the first groups to be classified based on taxonomy, on CBMs and on loop content. Recently, other types of GH19 chitinases

were characterized from Proteobacteria [127-130], while inactive plant isoforms of GH19 [131, 132] and a recent class I plant chitinase lacking loop 3 was described [120], raising doubts about the capacity of actual class-based GH19 classification to describe the observed sequence diversity.

Moreover, other three distantly related GH19 clusters (III, IV and V in [126]) have been identified based on sequences from bacteriophages, prophages and genomes of Proteobacteria and Nematoda. Among these, cluster III comprises the emerging group of enzymes characterized as endolysins with lysozyme-like properties [112, 113, 133-138], among which one 3D structure was solved (**Fig. 1.9B**) [113].

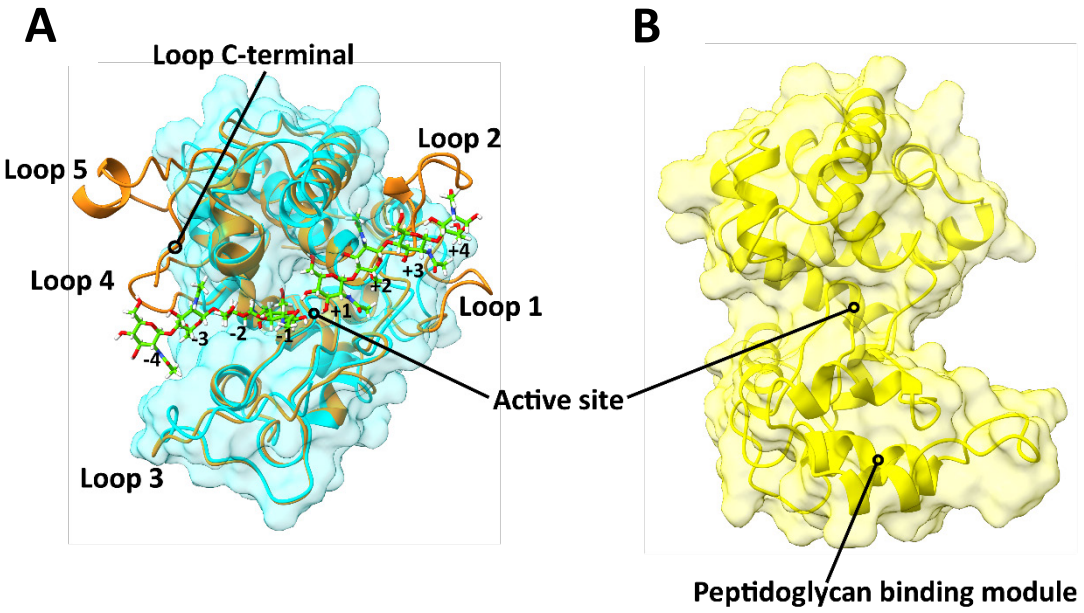


Fig.1.9 In **A**, the structures of a GH19 "loopful" chitinase from the rye seed *Secale cereale* (orange, PDB code: 4jol [139]) and a "loopless" chitinase from the moss *Gemmabryum coronatum* (cyan, PDB code: 3wh1 [111]) are superposed, showing the five additional loops of a "loopful" plant chitinases (only loop 3 is present in both). The two chitotetrasaccharides spanning the catalytic cleft in complex with the crystal structure of rye seed are shown. The numbers under the moieties of the chitotetrasaccharides are in accordance with the standard nomenclature for GH: cleavage occurs between units bound in subsites -1 and +1 [140]. In **B**, the structure of a GH19 endolysin from the bacteriophage SPN1S (PDB code: 4ok7 [141]) of *Salmonella typhimurium*.

1.4.3 Biotechnological applications

GH19s play an important role in plants as pathogenesis-related proteins overexpressed in defense against chitin-containing pests (like Fungi and insects) [117, 142-147]; an increased pest tolerance was also demonstrated by *in vivo* experiments in transgenic plants overexpressing heterologous GH19 genes.

The main application of GH19s is related to their biological role in plants: they are considered valuable bio-control agents for treating crops against various types of pests.

Other industrial biotechnological applications of GH19s are the same as for the chitin hydrolyzing enzymes from other GH families: the produced chitooligosaccharides (COSs) are compounds proven to possess multiple properties, like probiotic [148] and anti-inflammatory activity agents in human gut [149, 150], and anticancer capacities [151]; on the other hand, chitinase can be directly used in combination with antifungal drugs for improved therapeutic treatment of human fungal infections [152].

In human economy, chitinases have been tested as valuable enzymes in the process of circular bioremediation of shellfish waste from sea food industry, after the required chemical extraction of its chitin content [153], that can be used for the production of COSs and as raw material to fed microorganisms in the production of single cell proteins [103, 154].

In a recent study, a mutated GH19 chitinase was developed to catalyze the chemo-enzymatic synthesis of chitotetraose [155]. Chitinases applications are summarized in **Fig. 1.10**.

In the last decade, interest is growing for the application of endolysins as specific antimicrobial agents (named also “enzybiotics”) towards Gram positive bacteria in the frame of the quest for new antibacterial drugs to fight drug resistance [156]. Recent developments raised expectations also for their possible use against Gram negative bacteria [157-159].

GH19 were demonstrated to be phage endolysins and, therefore, can provide new enzybiotics. An example is a recent discovery of an enzyme from this family with

outer-membrane permeabilizing capacity against hospital Gram negative bacteria strains [134]. The attention toward the GH19 potentialities in the antimicrobial biomedical field is expected to increase in the next future.

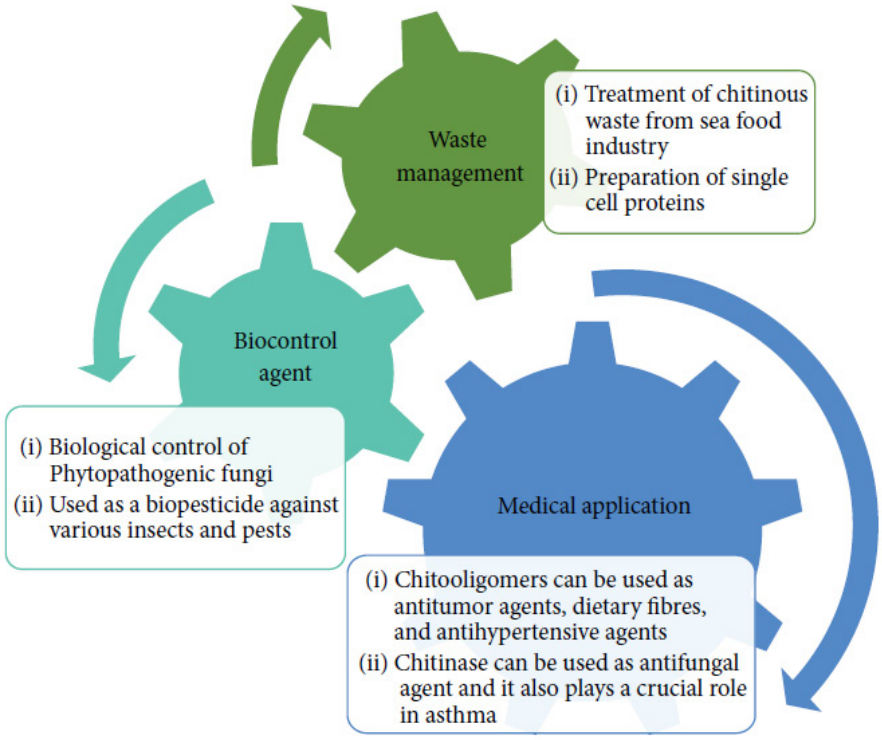


Fig. 1.10 Main applications of chitinolytic enzymes. Reproduced from [160].

1.5 Life in cold environments

1.5.1 Psychrophiles

Constantly cold environments are characterized by temperatures below 5°C and represent approximately 80% of the Earth biosphere, including polar regions, deep oceanic sediments and mountain glaciers [161, 162].

These environments are inhabited by psychrophilic and psychrotolerant (or psychrotrophic) organisms, the first ones growing well at temperatures around the freezing point of water and with an optimum below 15°C, the second ones having the capacity to survive around 5°C, but usually with an optimum of growth at mild temperatures [163]. Nevertheless, there is a continuum in temperature adaptation for life, with wide or narrow growth temperature ranges depending on the organism [161].

These organisms have to cope with many different challenges: first, any decrease in temperature exponentially reduce the rate of biochemical reactions; second, the viscosity of aqueous environments, which increases by a factor higher than two between 37 °C and 0 °C; third the increased production of reactive oxygen species induced by higher gas solubility at low temperature; forth the deleterious effects of cold on physical properties and functions of membranes, typically caused by a reduction in their fluidity, and a general impaired folding of proteins [162].

At temperatures close or below the freezing point of water there is also the problem to avoid the intracellular or extracellular formation of ice-crystals, which, in turn, cause direct and indirect damages to all biological membranes, causing cell lysis [164].

In order to cope with these problems, these organisms have evolved several strategies, including the modification of membranes composition to change their fluidity [165], the over-expression of anti-oxidative enzymes [166], the production of cold-shock proteins, the secretion and accumulation of cryoprotectant osmolytes and anti-freeze proteins (like ice-binding proteins [162]), and the production of cold-active enzymes [161].

1.5.2 Cold-active enzymes (CAEs)

In constantly cold environments, chemical reaction rates decrease exponentially with decreasing temperature, according to the Arrhenius equation for the catalytic constant (K_{cat}) [167] :

$$K_{cat} = Ae^{-\frac{E_a}{RT}}$$

Where E_a is the activation energy of the reaction, R is the gas constant, T is the temperature and A a collision frequency factor.

Cold active enzymes (CAEs) have the property to retain a significant fraction of activity at low temperatures with respect to their temperature of optimum, despite this phenomenon.

As suggested by the above equation, a possible solution CAEs adopt is the capacity to decrease the E_a required by the reaction more than what happens in mesophilic and thermophilic counterparts [168-177]. In this way, the K_{cat} of CAEs at low temperatures match more and less those observed for mesophilic enzymes at warm temperatures.

In order to better understand how this can happen, it is informative to look at the relation between the free energy of activation ($\Delta G^{\ddagger}$) between the ground and activated states, which represent the limiting step in a hypothetical reaction model. The enthalpic ($\Delta H^{\ddagger}$) and entropic ($\Delta S^{\ddagger}$) contributions are particularly important of $\Delta G^{\ddagger}$, which figures in Eyring equation for the K_{cat} calculation [178] :

$$K_{cat} = \frac{K_B T}{h} e^{\frac{-\Delta G^{\ddagger}}{RT}} = \frac{K_B T}{h} e^{\left(\frac{\Delta S^{\ddagger}}{R} - \frac{\Delta H^{\ddagger}}{RT}\right)}$$

Where K_B is the Boltzmann constant and h is the Planck constant.

To keep high level of K_{cat} at decreasing T, it is necessary to decrease $\Delta H^{\ddagger}$or increase $\Delta S^{\ddagger}$ for the reaction [179].

From a thermodynamic point of view, a decrease in $\Delta H^{\ddagger}$ can be obtained by a decrease of the number of interactions between the active site and the ligand that must be broken in the transition from the ground to the activated state [180]. The

increase in $\Delta S^{\ddagger}$ can be obtained by an increase in the flexibility of regions of the enzyme active site or of other regions on the protein surface [181]. This hypothesis was corroborated by recent computational studies [182].

This increased flexibility might drive a trade-off between activity and stability, so that generally a low stability is observed in CAEs (**Fig. 1.11**) [172, 183].

Side effects of the enhanced flexibility can be a looser binding of the substrate (higher Michaelis-Menten constant values), and an increase in the conformational space explored at the active site, whose flexibility might lead to promiscuity (capacity to bind and catalyse the same reaction on different substrates) [184].

By comparing corresponding sequence positions and structures of CAEs with mesophilic/thermophilic homologues, common trends to achieve the above mentioned thermodynamic changes include: a reduction of the number of ion pairs, disulphide bonds, hydrogen bonds and hydrophobic interactions; a decrease of subunit interactions for multimeric enzymes; an increased interaction with the solvent by reduced hydrophobicity in the core; an higher accessibility to the active site; a decreased cofactor binding; a clustering of glycine residues, and a lower proline and arginine content [185].

Besides the relevance of understanding the molecular and evolutionary grounds of cold activity, enzymes from psychrophilic organisms may find application in biocatalysis [183]. Because of their high activity at low temperatures, CAEs can help to reduce energy consumption and the environmental impact of biotransformation reactions, by avoiding side chain reactions for example in the food industry and in fine chemistry [186]. Moreover, operating temperatures are permissive for heat-labile and perishable substrates and raw materials. Not least, the possibility of inactivating CAEs by moderate heating can also be advantageous whenever the catalyst has to be removed at the end of a process [187].

Thus, CAEs can be used to re-design existing processes based on mesophilic enzymes or to develop new ones, promoting an up-coming “cold revolution” in different fields [183, 187, 188].

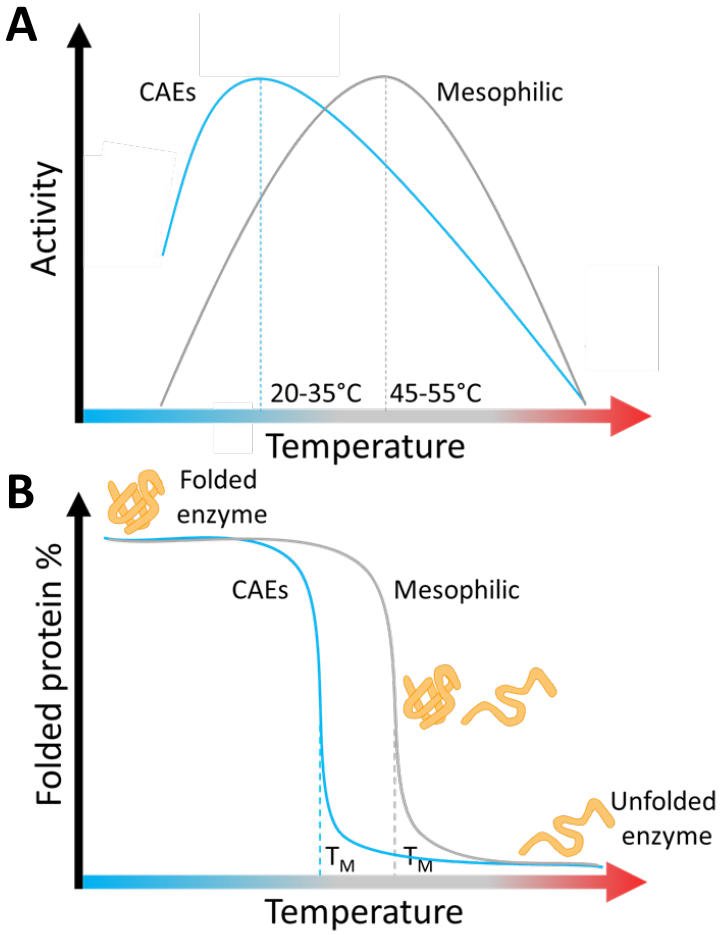


Fig. 1.11 Activity and heat-driven unfolding of psychrophilic enzymes. **A.** Temperature dependence of a generic CAE. Generally, CAEs exhibit optimal temperature for catalysis (T_{opt}) in the range from 20 to 30°C and maintain a relatively high activity at low temperature. **B.** The inactivation of psychrophilic enzymes usually anticipates the temperature at which there is the loss of protein structure (T_M, temperature of melting), suggesting that the thermolability concerns first their active sites. Adapted from [183].

1.6 *Pseudomonas sp* strain Ef1 from an Antarctic bacterial consortium

Euplotes focardii is an unicellular free-swimming ciliate endemic of the oligotrophic coastal sediments of Terra Nova Bay, Antarctica [189]. It has an optimal growth temperature around 4°C and its viability decreases upon exposition to temperatures above 10°C [190]. It was collected from expeditions in Antarctica and then stored at low temperature in laboratory conditions since 1990, in order to use it as a model for the study of biochemical and physiological processes in the cold [191]. It was proposed that this ciliate lives in association with a consortium of bacterial symbionts that were found in its recently analysed metagenome [191].

Among these Bacteria, a *Pseudomonas sp.* strain Ef1 was isolated and its genome sequenced (deposited in GenBank under the accession number VAUR00000000) [192]. Two GH19 sequences coding for putative chitinases, LYS177 (Uniprot accession: A0A516Z9W0) and LYS188 (Uniprot accession: A0A516Z9V1) were automatically annotated by Prokka [34].

Reported in the first part of this thesis is the biochemical and biophysical characterization of the first two GH19 enzymes to come from a psychrophilic organism. Moreover, as GH19 lysozymes and chitinases have potential applications in human health and nutrition, in the second part of this thesis the analysis of the GH19 sequence space is reported.

This study has contributed to the description of endolysins with cold-active features, to the identification of sequence signatures for predicting substrate specificity in GH19, and to the analysis of observed biochemical, sequence and structural diversity in this family, relevant to the identification of GH19 enzymes with interesting new features.

2.Main results and discussion

Although sharing the same fold and catalytic residues, GH19s are specialized as chitinases and endolysins. The activity, structure and biological function of GH19 chitinases have been experimentally characterized in plants and Actinobacteria since decades [125, 154], while few studies have described GH19 endolysins to date.

The exploration of biodiversity inhabiting extreme environments could allow the discovery of GH19 enzymes presenting features never described to date, considering that no extremozyme has been characterized to date from this family.

In a first work, two single domain GH19s, LYS177 and LYS188, were identified in the genome of a *Pseudomonas* sp. Ef1, a bacterium [192] living associated with the marine Antarctic ciliate *E. focardii* [191]. A manuscript is currently in preparation and a draft which includes also the methods is reported in **chapter 3** of this thesis.

The far-UV circular dichroism spectra (190-260 nm) confirmed that both enzymes have a globular structure mainly made of alfa helices, typical of the GH19 fold, as can be appreciated by looking to models built by homology modelling (**Fig. 2.1A**-**B**). Despite most of the characterized GH19s display chitinolytic activity, the phylogenetic analyses (**Fig. 2.2**) show that LYS177 and LYS188 are closely related to enzymes characterized as endolysins [112, 113, 193]. Moreover, by looking to the genomic context in the natural *Pseudomonas* host, both enzymes were found in prophagic regions.

The consequent hypothesis that both enzymes are endolysins was confirmed experimentally as the two heterologous purified enzymes were not active in the hydrolysis of insoluble and soluble chitinase substrates (chitin azure and a chitooligomeric synthetic chromogenic analogue), while proved to be able to lyse *Micrococcus lysodeikticus* (Gram positive bacterium) cells. By performing the same lysozyme assay in different conditions of temperature (**Fig. 2.1C**-**D**), the obtained results permit to show that LYS177 and LYS188 fit the canonical definition of cold-active enzymes, since they retain a relatively high activity at low temperature (≈ 40%). Moreover, the temperature optimum of LYS177 and LYS188 were around 20°C and 30°C, respectively, lower than the estimated temperatures of midpoint denaturation, around 50°C and 45°C (**Fig. 2.1E**-**F**). These results are in agreement with an active site

having high conformational flexibility with respect to the overall structure, a feature used to explain how a relatively high catalytic rate can be maintained at low temperature [172, 184].

Experiments of incubation of LYS177 at different temperatures (4°C, 20°C and 37°C) and increasing time resulted in retaining of activity and secondary structure only at 4°C, while at higher temperatures a parallel loss of both was measured after few days. Overall, this result confirmed that LYS177 is thermolabile. Even if this is considered a typical feature of cold active enzymes [182], the issue of CAEs thermolability is far to be fully unveiled, with enzymes losing their activity within few hours only at mild temperatures [194], to cases of thermostability similar to that of mesophilic or even thermophilic homologues [195].

Inspection of the LYS177 and LYS188 sequences aligned with other characterized endolysins showed that these enzymes lack two helices of the peptidoglycan binding 3-helix bundle (named in this study “peptidoglycan binding module”, PBM, shown in **Fig. 1.9**), present in the GH19 endolysin of *Salmonella typhimurium* bacteriophage SPN1S, from which a 3D structure is available (PDB 4ok7). In [141] this region was experimentally shown to increase the affinity for the cell wall of the host.

The high variability observed in sequence composition and length of the PBM region also in other characterized GH19 endolysins led to the hypothesis that it could be important for co-evolutionary phage-host interactions, considering the variability in the cell wall composition of bacteria [196]; anyway, not enough endolysin sequences were analyzed in this study to further support this hypothesis.

This work focused on the properties of two enzymes described as cold-active lysozyme endolysins; thus, conservation analyses were not performed on sequences of chitinases, and only characterized GH19 proteins (75 sequences) were considered to build the phylogeny. Nevertheless, average pairwise distance between all identified GH19 subgroups was calculated to be 60,1 ± 7,3%, indicating that this family most likely has a very remote common ancestor and its diversity is potentially underestimated to date.

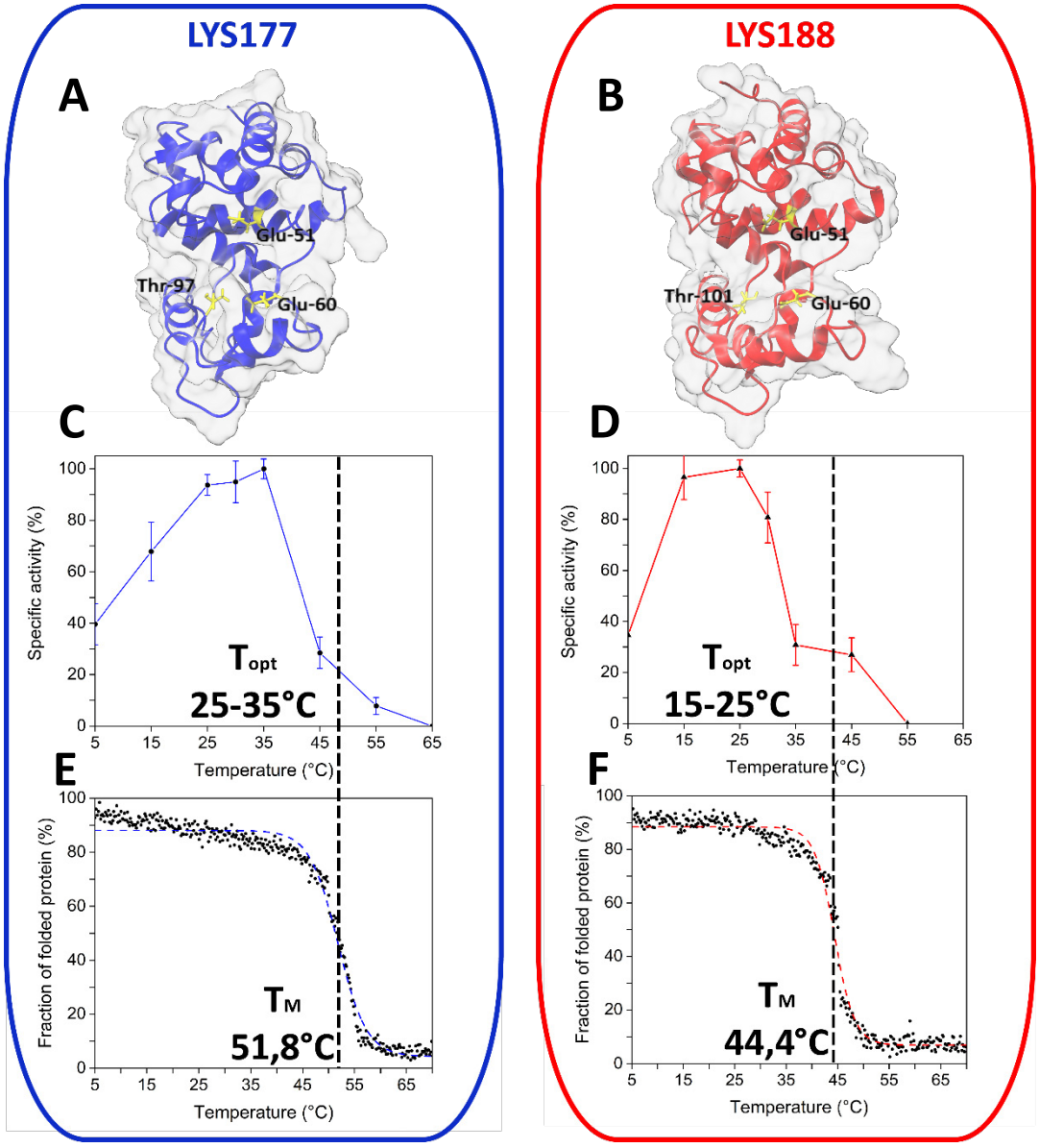


Fig. 2.1 In **A** and **B** the 3D models, built by RaptorX server [197], of LYS177 and LYS188 are displayed in cartoon and transparent surface, with the catalytic triad highlighted in yellow. In **C**-**D** and **E**-**F** the specific activity at pH optimum and the relative intensity of CD signal at 222 nm (used as a proxy for the fraction of folded protein) are plotted at different temperatures for both LYS177 and LYS188. The scatterplots in **E** and **F** were fitted with a Boltzmann distribution. T_{opt} = optimum temperature. T_m = midpoint denaturation temperature.

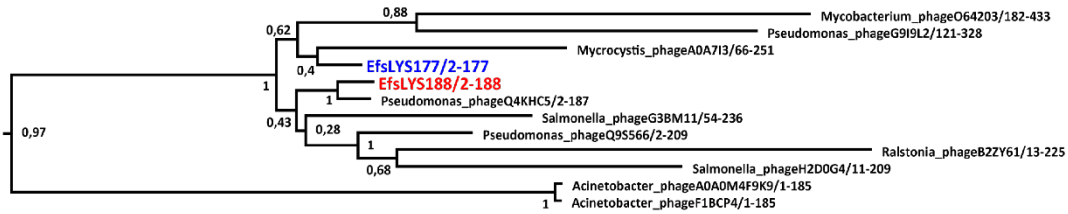


Fig. 2.2 Supported monophyletic clade including all characterized endolysins to date, including LYS177 and LYS188. This clade was extracted from the phylogenetic analysis of the 75 characterized GH19 sequences, reported in **Chapter 3** of this thesis. LYS177 is named EfsLYS177, highlighted in blue, and LYS188 is named EfsLYS188, highlighted in red.

Therefore, a study on GH19 diversity in sequence and structure, and its evolution, is required to have a more complete picture on GH19 properties. This is relevant also for the potential applications of GH19s in the field of human health and crop treatments against pests (see **paragraph 1.4.3** of this thesis).

GH19 evolution has been studied by previous authors [114, 126], but mainly GH19 chitinases from plants and Actinobacteria were sampled [89, 103, 104, 198, 199]. Moreover, the substrate specificity of chitinases seem to be high, as demonstrated in works in which substrate promiscuity was tested [137, 200-205]: weak or no lysozyme activity was detected for GH19 chitinase. Thus, this indicates also the importance to identify sequence signatures associated with substrate specificity for understanding the molecular basis of substrate specialization and for further improving the activity of GH19 enzymes.

In order to answer to these requests, a bioinformatic investigation of the protein sequence space, a recently applied approach in the field of enzyme discovery [93, 94, 206], was performed for GH19 family, and its evolution explored by phylogenetic analyses. The main findings of this work will be presented and discussed below; a manuscript on this work has been submitted to *The FEBS Journal* and the draft is included in **chapter 3** of this thesis, together with the details on methods employed to achieve the results.

The starting point has been the creation of a database named GH19 engineering database (GH19ED), containing 22461 protein sequences from NCBI non redundant

and Protein Data Bank public databases, with at least a GH19 domain. All the GH19 sequences from this database were clustered to pick up 5229 centroids (i.e., a sub-sample of all the sequences, that is representative of the sequence space) and generate similarity networks with them. By applying a 40% identity cut-off to the edges connecting the centroids (the nodes of the network), more than 90% of the sequences are divided into two big clusters (8554 and 10967 sequences) containing, respectively, the sequences of characterized chitinases and endolysins (**Fig. 2.3**). These GH19 clusters were used to define two subfamilies (named "superfamilies" in the submitted draft, for compatibility with the ontologies used in the database), one containing putative chitinases (CHITs) and the other putative endolysins (ELYSs). Sub-clusters within each subfamily were also obtained by applying a 60% identity cut-off, which permitted to split CHITs sequences into 18 groups, used to define 17 homologous families that are coherent with previously defined chitinase classes [117, 119]. Two of these include most of the characterized GH19 (49 out of 75) divided in plant "loopful" (class I and II) and "loopless" chitinases (class IV); the latter group is a merge of two sub-clusters that both contain sequences characterized as class IV "loopless" chitinases from plant or from Bryophyta. Therefore, the system of classification built in this study takes also into consideration the recently introduced difference between "loopful" and "loopless" chitinases [111, 207], used to indicate the presence or absence, respectively, of different combinations of six loops around the catalytic cleft [111, 201]. Among other homologous families defined in this study, two are new groups containing plant sequences: it is likely they contain non-active chitinase like proteins (CLPs), as the sequences in this groups were characterized as lectins or mediators of plant physiological responses to environmental conditions.

CHIT sequences from organisms other than plants were divided in eight homologous families from various bacterial taxa, included the most studied bacterial "loopless" chitinases that come from Actinobacteria fir > 90% of sequences [125]. This homologous family includes also minor fractions from Myxococcales (> 3%), Firmicutes (> 1%), Betaproteobacteria (> 1%) and Gammaproteobacteria (> 1%),

enriched in species typically found in soils. Other five small clusters are from other Eukaryotes (Oomycota, Fungi and Metazoa).

ELYSs sequences were divided in 34 homologous families from phage/bacterial sources. Only two predominant homologous families contain each more than 2500 sequences, while there are 26 homologous families of small size (from few to 100 sequences), more than double than in CHITs. Therefore, the sequence space of GH19 seems to be rather connected for ELYSs with respect to CHITs. This may be due to a bias in sequence sample available in public databases and not a real difference in terms of distribution of diversity in the sequence space.

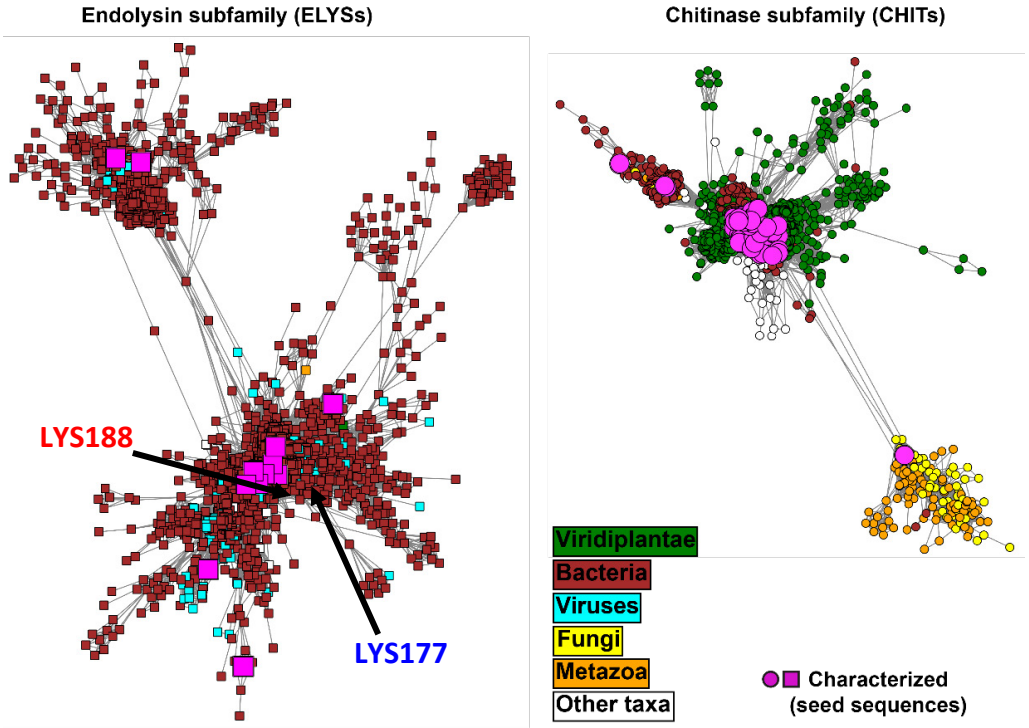


Fig. 2.3 Protein sequence networks of representative domains of the two bigger clusters containing characterized sequences (5067 centroid nodes in total, 2738 nodes on the left for ELYSs and 2329 nodes on the right for CHITs) connected by edges with 40% minimum global identity cut-off. The two endolysins characterized in the first part of this thesis are indicated by black arrows. The prefuse force-directed OpenCL layout of Cytoscape 3.7.1 [208] was used. The domains were extracted from Pfam GH19 Hidden Markov Model profile (PF00182) scanning of the sequences collected through BLAST searches using the characterized sequences as queries. Nodes are coloured according to their taxonomic source annotated in NCBI.

Overall, an extended and biochemically integrated classification system for all the GH19 sequences in CHIT and in the under-characterized ELYS subfamily was created, providing a scaffold that will be populated after more experimental information will be available, to better predict the existence of interesting functional variants occupying unexplored portions (nearby or far from any characterized protein) of the sequence space in each subfamily, especially for ELYS.

Since there is a clear distinction between putative endolysins contained in ELYSs and putative chitinases in CHITs by looking to sequence similarity networks, conservation analyses were performed on ≈ 300 centroids from ELYs and CHITs, to identify diagnostic sequence positions to serve for immediate prediction of subfamily specificity.

The results, plotted on reference structures of CHITs and ELYSs (**Fig. 2.4A**-**B**-**C**-**D**), revealed that all GH19 possess a conserved core made of 27 residues (**Fig. 2.4E**-**G**) that comprise the central portion of the active site cleft binding the substrate at subsites -2, -1 and +1. The three catalytic residues part of this core are conserved and rarely substituted, in particular the catalytic acid, which is typically replaced in GH19 proteins that play a non-enzymatic function, such as lectins or physiological pathway regulators (as demonstrated in [209-212]). Many structural and kinetic studies involving GH19 chitinases support the hypothesis that these central subsite positions are involved in positioning the substrate in place for starting the hydrolysis [111, 125, 139, 201, 213, 214]. Also other hybrid experimental/theoretical kinetic model based works converged towards this result [215, 216].

On the contrary, sequence plasticity was detected at both extremities and on surface elements around the substrate binding cleft, comprising the CHITs loop motifs (**Fig. 2.4B**) and the PBM of ELYSs (**Fig. 2.4D**). Indeed, CHITs loops are overall not conserved in their primary sequence, with exception of loops 3 and 4, and vary in length, with exception of loops 4 and 5, the conclusion is that most likely they are flexible motifs responsible for accessory properties and are not necessary for the catalysis. This is also confirmed by experimental studies on CHIT loops, which conclude that they are flexible structures [121, 139] that can significantly increase

the size of the binding cleft, altering the binding mode of COSs [111, 121, 125, 139, 201, 217-220] and reducing in most of the cases the catalytic efficiency on short soluble substrates [122, 201, 207, 216, 218, 221].

The PBM alone was demonstrated to have an *in vitro* affinity for peptidoglycans higher than the enzyme after its removal [141]; the low level of conservation detected for this module in the results of this thesis is coherent with the hypothesis reported in the first part of this thesis: it could be a region regulating the adsorption of the enzyme on the cell wall, useful for the phage to cope with the peptidoglycan composition variability of the bacterial host [196].

Overall, the performed conservation analysis leads to the conclusion that GH19 share a conserved a core essential for the binding of GlcNAc containing polymers: this is not surprising because some CHITs and a single ELYS can bind and hydrolyze murein and chitin, respectively, and these substrates are chemically similar.

In [141], a GH19 endolysin structure was superposed to structures of enzymes from other families in the lysozyme superfamily, concluding that it is functionally an N-acetyl-β-D muramidase because the position of catalytic acid and base residues is compatible to C-type lysozyme (GH22). However, GH22 enzymes have a retaining catalytic mechanism with the involvement of acetamido substituent at C2 of the substrate during the catalysis [106]. Instead, we suggest the mechanism of hydrolysis of GH19 ELYSs to be similar to the GH19 CHITs. Anyway, the superposition of ELYS and CHIT reference structures suggests that ELYSs differ from CHITs by a larger substrate binding cleft at subsites from -4 to +3, to accommodate the more bulky murein substrate, and possibly by a different opening or closing dynamic of the cleft during catalysis. In this regards, six positions in the generated alignments were specifically found in CHITs, while absent in ELYSs (and, *viceversa,* four positions in ELYS, while absent in CHITs), and defined as hallmarks of activity specialization. These residues (**Fig. 2.4F**-**H**) are located apart from the catalytic cleft and do not interact directly with the substrate, on the hinge or inside the lobes of the reference structures, with a possible role in the control of lobe flexibility during the reaction; chemical shift perturbations were experimentally observed behind the binding groove of GH19

"loopless" chitinase from *Gemmabryum coronatum* [111], and domain motion was often reported to be required upon the ligand binding in bi-lobal glycosyl hydrolases [222]. However, it is not possible to demonstrate if these signatures effectively contribute to the molecular mechanism responsible for the observed substrate specificity. Nevertheless, the amino acid pattern observed for these signatures was tested on characterized GH19: 62/63 chitinases possess at least 4/6 positions of CHIT signature, and 10/12 endolysins possess at least 3 /4 positions of ELYS signature. Therefore, the signatures are general enough to be used for an immediate distinction between GH19 subfamilies.

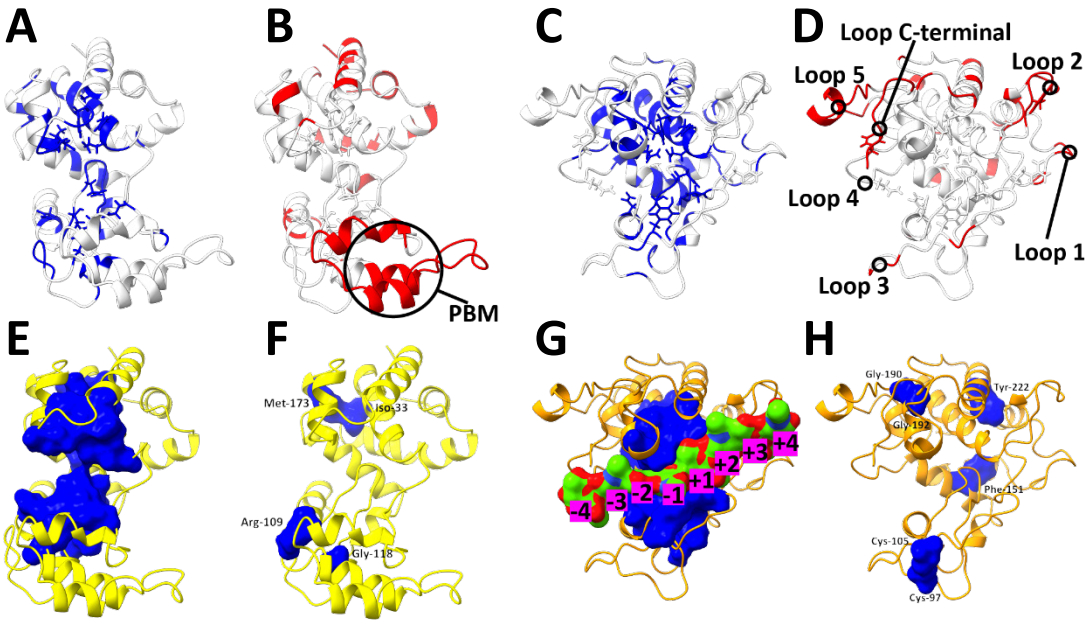


Fig. 2.4 Least conserved (red) and most conserved (blue) residues plotted onto (**A**-**B**) the ELYS reference model from bacteriophage SPN1S (PDB code: 4ok7), and (**C**-**D**) the CHIT reference model from rye seed (PDB code: 4j0l). The models are visualized in cartoon, with substrate binding residues as sticks (except glycine). (**E**-**G**) The most conserved and structurally aligned positions between CHITs and ELYSs are plotted as blue solvent accessible surface on the same models. (**G**) The surface representation of two chitotetrasaccharide molecules spanning the substrate binding cleft of the rye seed chitinase model is visualized and the subsite numbers are given below. The residues corresponding to ELYS (**F**) and CHIT (**H**) sequence signatures identified in this work are shown as blue solvent accessible surfaces and labelled.
PBM = Peptidoglycan binding module.

Interestingly, The only CHIT outlier is a GH19 chitinase characterized from a fungus species [223], located far from the other characterized CHITs, in the lower-right portion of the CHIT network in **Fig. 2.3**, and the two characterized endolysins not fitting the ELYS signature are located apart from others, on top left portion of the ELYS network in **Fig.2.3**: these sequences come from the phages of two *Acinetobacter* species in which they were reported to possess an amphipathic helix inducing outer membrane permeabilization [112, 134]. Therefore, the signatures could also be useful for identifying outliers with possibly novel properties. As no variants have been studied at these positions, yet, next studies are encouraged to focus on the functional role of these positions by employing site-directed mutagenesis and simulations.

The evolution of structural motifs (chitinase loops, endolysin PBM) and accessory binding modules was investigated from an evolutionary point of view, by building a molecular phylogenetic tree of centroids representative of all homologous families.

Based on the results of this analysis (**Fig. 2.5**), CHITs loops 3 and 4 were most likely the first to be present in the common ancestral CHIT sequence, followed by additions of loops 1 and 2 in the plant lineage, and of loops 5 and C-terminal in the plant "loopful" lineage (IDs 1-3-4), after its separation from "plant loopless" lineage (ID 2A-B), which lost loop 4. This genealogy can explain why plant "loopless" chitinases are structurally similar to bacterial "loopless" chitinases (ID 5) and do not share the loop content of plant "loopful" chitinases: the "loopless" and "loopful" chitinases probably evolved from two different ancestors before the appearance of a full set of loops. In *Urtica dioica* like CLP lectins (ID 4) loop 5 was secondary lost. "Loopful" plant chitinases passed to bacteria by two HGTs and lost loop 1 in both cases; some other loops also disappeared in three bacterial taxa (IDs 7-10-11) and in Proteobacteria chitinase (ID 6) loops 2 and 3 became longer.

Overall, during GH19 evolution, up to four different loops were acquired in CHITs, with only the most conserved loops 3 and 4 likely present since the chitinase common ancestor, while some were secondary lost in different organisms: this permitted GH19 to explore different combinations of properties, still unknown, potentially optimized for slightly different functions in different taxa. As just three GH19 chitinases were

described in groups of organisms other than plants and Actinobacteria, other bacterial taxa would require more attention in future experimental studies. This is particular true for members of homologous family 7, which have a modified N-terminal region without the first three loops, and the homologous family 6, as loops 2 and 3 became longer, and evidence of exo-activity was experimentally collected in the two characterized GH19 chitinases of this group, from *Vibrio proteolyticus* and *Pseudoalteromonas tunicata* [129, 130]. GH19 CHITs are typically endo-acting enzymes. Thus, the appearance of longer loops during the evolution of homologous family 6 may explain the processive exo-activity, as supported by observations on other GHs, where processivity was considered as the result of longer loops responsible for the conversion of the active site shape from a cleft to a tunnel [95].

Different combinations of accessory modules, added to GH19 during evolution, exist. By looking to GH19 tree genealogy (**Fig. 2.5**), the most likely hypothesis to explain the distribution of CBM18, 5/12 and 13 during CHITs evolution is that CBM18 was added in the plant lineage (IDs from 1 to 4), but was not transferred to bacteria, as happened to the catalytic domain, when there were the two HGTs to homologous families with IDs 6-7-9-10, which likely received the CBM5/12 from bacterial “loopless” CHITs (ID 5). Some members of the bacterial “loopless” chitinases possess also a CBM13, a domain with loose and broad sugar specificities [99], that was found in association to xylanases (GH10) from Actinobacteria. It is likely that this is the most recent accessory domain that recombined with a GH19, as it is limited to a few sequences in the same homologous family. The same applies for LysM, which is a ubiquitous non-catalytic motif repeat that was shown to bind both peptidoglycan and chitin in plants [224], and was mainly found in a small subgroup of "loopless" bacterial CHITs from Cyanobacteria (ID 11).

The distribution of PBM insertion is restricted to just two ELYSs homologous families (IDs 2-19) and evolved recently as these groups share a common ancestor with only another homologous family (ID 30), which agrees with the hypothesis reported above.

In ELYSs, only two known accessory binding modules are present (PG_binding_1, and LysM) in a few hundred sequences (represented by more than 10 centroid

sequences): they are spread among different homologous families with low sequence similarity. This finding supports the hypothesis of multiple independent recombination events with the ELYS catalytic domain.

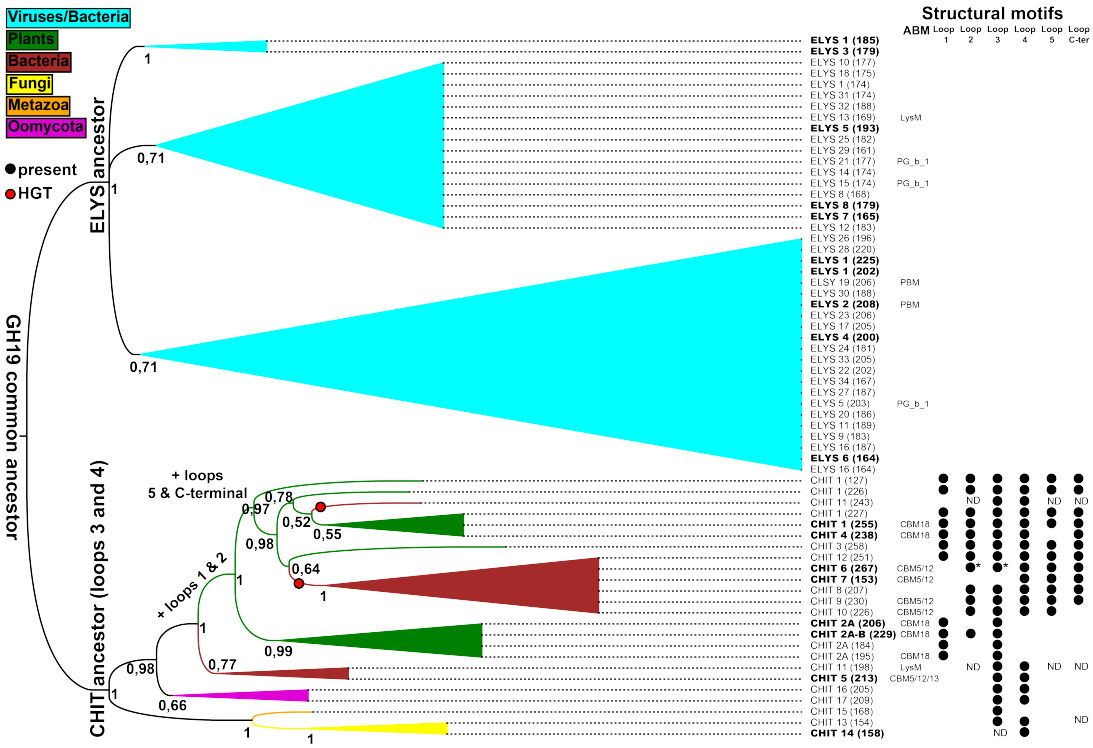


Fig. 2.5 Phylogeny of centroid domains representative of GH19 homologous families. Structural motifs and accessory binding modules analysed in this study for each homologous family are plotted. Sequences are indicated with the respective subfamily code, homologous family ID and sequence length in parentheses. The header of centroids representing groups with characterized sequences inside are in bold. Some of the nodes have been collapsed for a better visualization.

*These loops are longer than the length distribution of the corresponding loop in other homologous families.

HGT = horizontal gene transfer; ABM = accessory binding module; ND = not defined; ELYS 1 = *Pseudomonas* prophage like; ELYS 2 = *Salmonella typhimurium* like; ELYS 3 = *Salmonella* phage PVP-SE1 like; ELYS 4 = *Ralstonia* phage like; ELYS 5 = *Pseudomonas* phage OBP like; ELYS 6 = *Acinetobacter* phage like; ELYS 7 = *Mycrocystis* phage like; ELYS 8 = *Mycobacterium* phage like; ELYS 9 to 34 = other putative endolysins from phages and prophages; CHIT 1 = plant "loopful"; 2A-B = plant "loopless"; CHIT 3 = plant CLP with regulatory functions; CHIT 4 = *Urtica dioica* like CLP lectins; CHIT 5 = bacterial "loopless"; CHIT 6-7 = Proteobacteria; CHIT 8 to 12 = other putative chitinases from Bacteria; CHIT 13-14 = Fungi; CHIT 15 = Metazoa; CHIT 16-17 = Oomycota.

By analyzing the taxonomic distribution of GH19 during its evolution, it is possible to see that after an early separation between CHITs and ELYSs. The endolysin lineage remained confined in the genomes of phages and bacteria, while chitinases spread in both prokaryotic and eukaryotic taxa, in which they independently evolved up to the appearance of bacterial and plant lineages. The spread of "loopless" bacterial chitinases (ID 5) from soil bacteria can be associated to the advantage they provided to exploit the carbon source contained in chitinous organic matter and for inhibiting the growth of competing fungal species. Waves of enzymatically active GH19 over-expression were measured in plant tissues after exposure to various stresses, including mechanical wounding and infections by fungal pathogens [145, 146, 225, 226]. Therefore, these enzymes most probably spread in plants because they entered in the systemic acquired response mechanism activated against chitin-containing pathogens. More recently, plant homologous families with ID 3-4 evolved as CLPs, non-enzymatic coagulant factors in latex or regulators of plant growth in response to changing environmental conditions. This leads to the hypothesis that this functional transition from enzymatic to regulative activity may have been favored by the pre-existing over-expression in plant defense responses.

The evolutionary hypothesis reported in this study revealed an evolutive scenario different from the one previously described in [114], where plants were considered to be the original hosts of GH19 genes and secondary transferred to Actinobacteria. The molecular phylogenetic study reported in this thesis considered a sequence sample that is representative of the GH19 sequence space known so far, and not oversampled the most considered groups of plants and Actinobacteria. Moreover, a Bayesian approach with an underlined statistical model of molecular evolution was applied. Therefore, robust evidence was collected to support the hypothesis that bacterial and plant GH19 evolved independently and in parallel, while HGTs moved GH19 to bacteria just after the diversification of plant chitinases. A byproduct of this approach was that sequences from superphylum Alveolata and phylum Nematoda were not represented because in GH19ED database very few sequences belong to these taxa and the analysis was possible only on a representative sub-sample of all

sequences in the GH19ED. If more sequences from rare taxa will be added in the NCBI in the next future, another analysis would be necessary to complete the evolutive scenario of GH19s.

In conclusion, in this thesis two GH19 lysozyme endolysin from an Antarctic genome were characterized as cold-active and thermolabile enzymes, suggesting potentialities for future testing of LYS177 as an anti-bacterial agent to treat food/beverage stored at low temperature. These sequences were initially annotated as putative chitinases, probably because the currently available GH19 CAZy classification is limited. Therefore, the sequence space of the whole family was explored by a bioinformatic study, establishing the common and divergence features of chitinases and endolysins. Single position signatures suitable to predict activity specificity were provided, and the plasticity of GH19 sequence and structural diversity was embedded in its evolutionary history.

This information was integrated in the public accessible BioCatNet database system [227] (https://www.gh19ed.biocatnet.de), with a defined standard numbering scheme for each position of the sequences and a binary code to describe the presence/absence of structural motifs and their distribution. As the consequence, the content of this thesis will provide useful insights for future discovery of GH19 enzymes and for understanding the molecular mechanisms of substrate specificity in order to find activity enhancing mutations. The capacity to predict and exploit new GH19 functions and properties will benefit also from future efforts into biochemical and structural characterization of still unknown and unannotated accessory domains, especially for GH19 endolysins.

3.Drafts

Antimicrobial endolysins from Antarctic *Pseudomonas* display lysozyme activity at low temperature

Marco Orlando[1], Sandra Pucciarelli[2], Marina Lotti[1*]

[1] Department of Biotechnology and Biosciences, State University of Milano Bicocca, Milano, Italy

[2] School of Biosciences and Veterinary Medicine, University of Camerino, Camerino (MC), Italy

Correspondence to: Marina Lotti, Department of Biotechnology and Biosciences, State University of Milano Bicocca, Piazza della Scienza 2, 20126 Milano, Italy

Tel: ++39 02 64483527; marina.lotti@unimib.it

Author's ORCID

Marco Orlando: 0000-0002-5914-3052

Sandra Pucciarelli: 0000-0003-4178-2689

Marina Lotti: 0000-0001-5419-7572

Abstract

Organisms adapted to thrive in cold environments produce enzymes with the remarkable ability to perform catalysis even at temperatures approaching the freezing point of water. Such cold-active proteins show adaptive changes when compared with mesophilic homologues and are of interest for several processes where low-temperature activity coupled with easy thermal inactivation can be of advantage. We have identified and characterized two glycosyl hydrolases of family GH19 in the genome of an Antarctic *Pseudomonas* strain, member of the microbial consortium associated to the psychrophilic ciliate *Euplotes focardii*. Both recombinant proteins showed optimal temperatures of about 25-35°C and were able to retain 40% of their highest activity at 5°C, thus conforming to the definition of cold-active enzymes. Based on sequence analysis and on activity assays, we hypothesize they are specialized phage endolysins with lysozyme activity integrated in prophagic regions of the *Pseudomonas* host. The best performing of the two, named LYS177, is active and stable over several days at 4°C and displays activity on both Gram-positive and Gram-negative bacteria.

Keywords: cold-adaptation, cold-active enzyme, endolysin, glycoside hydrolase 19, antibacterial activity

Introduction

Adaptation to life in cold environments translates to a variety of molecular changes in the cell structures and in the macromolecules of the so-called psychrophilic organisms [161, 162]. It is interesting to observe that the body of information built over the years shows a large diversification in adaptation strategies that highlights, beside the relevance of the selective pressures exerted by temperature, also the importance of the evolutionary history organisms followed. This is obvious, if one compares sequences and biochemical features of cold-active enzymes, where cold activity is defined as the ability to retain relevant residual activity at temperatures close to 0°C. It is broadly accepted that most cold active proteins can cope with catalysis at low temperature since they are endowed with high flexibility either in the whole structure or at least in the regions surrounding the enzyme active site [180, 228]. This property may reflect in a decrease of thermal stability of the whole structure or in localized protein lability. The latter often results in the loss of enzyme activity at a temperature lower than the temperature of melting, because of the uncoupling between overall structural denaturation and inactivation [172, 229].

Besides the relevance of understanding the molecular and evolutionary grounds of cold adaptation, enzymes from psychrophilic organisms may find application in biocatalysis [185]. Indeed, high specific activity at low temperature is of advantage for energy saving and in processes in which heat sensitive substrates are used and side reactions should be avoided, as for example in the food industry and in fine chemistry. A further benefit of psychrophilic heat labile enzymes may rely on the possibility to inactivate them through small temperature increases [185, 230].

We are studying Antarctic bacteria and unicellular eukaryotes sampled at Terranova Bay as sources of cold-adapted enzymes and antifreeze proteins [166, 231]. This work focuses on glycoside hydrolases belonging to family 19 (GH19), according to the classification of CAZy, the database of carbohydrate active enzymes [79]. GH19 enzymes are endo-glycosydases that hydrolyse β-1,4 glycosidic bonds by inverting the anomeric configuration of the C1 [110]. They are classified either as endochitinases (EC: 3.2.1.14) that cleave glycosidic bonds between N-acetyl-glucosamine (GlcNAc) residues within chitin chains, or as lysozymes (EC: 3.2.1.17) cutting peptidoglycans between N-acetylmuramic acid and GlcNAc residues.

Earlier reports included in GH19 plant chitinases only [117-119]. Later on, bacterial chitinases were added and events of horizontal gene transfer where hypothesized to account for their occurrence in this family [114, 126]. More recently, a number of papers [232, 233] pointed out that GH19 proteins are coded by single genes or modular multi-domain lytic gene cassettes in bacteriophages or prophages. Some of these enzymes have been biochemically characterized and classified as endolysins with lysozyme-like properties [112, 113, 133, 136, 137, 193]. Endolysins are produced in the late phases of the bacteriophages lytic cycle and allow for the release of the phage progeny since they attack the peptidoglycan polymer of the host cell wall. Endolysins are under investigation as specific antimicrobial candidates towards Gram-positive bacteria in the frame of the quest for new antibacterial drugs to fight drug resistance [156]. Recent developments raised expectations also for a possible use against Gram-negative bacteria [157-159].

In this study we report about the features and the evolution of two GH19 endolysins identified in the genome of the Antarctic *Pseudomonas* Ef1 bacterium, isolated from a microbial consortium associated

to the strict psychrophilic protozoan *Euplotes focardii* [191]. At the best of our knowledge, this is the first work to report the functional characterization of psychrophilic lysozyme-type endolysins. *Pseudomonas* Ef1 endolysins display high murein hydrolase activity at 5°C and are inactivated at mild temperature, making them candidates for future testing as thermolabile long-term antimicrobial enhancers during beverage and foodstuff fridge storage.

Materials and Methods

Identification of the bacterial strain and of glycoside hydrolases sequences

Pseudomonas Ef1 is a Gram-negative bacterium that was isolated from the microbial consortium previously described as associated to the psychrophilic Antarctic ciliate *Euplotes focardii* [191].
In order to assess the taxonomic position of this strain with respect to other *Pseudomonas* species, the 16S rRNA gene was amplified from the genomic DNA (Ramasamy et al**.,** Microbial Resource Announcement, submitted) by PCR using bacterial universal degenerate primers 27F (5'-AGAGTTTGATCMTGGCTCAG 3') and 1492R (5'-TACGGYTACCTTGTTACGACTT 3'), as forward and reverse primers, respectively. Amplification was in a Biometra Thermal Cycler (Biometra Ltd., Kent, UK) with the following cycling conditions: initial denaturation at 94 °C for 5 min, 30 cycles of 1 min denaturation at 94 °C, annealing at 60 °C for 1 min, and extension at 72 °C for 1 min. A final extension step was at 72 °C for 5 min. Sanger sequencing of the 16S rRNA amplicon was performed by BMR Genomics (Padova, Italy). The rRNA sequence was used as query for a Blastn search on the NCBI data bank (http://blast.ncbi.nlm.nih.gov).
Glycoside hydrolase sequences were identified in the genome deposited in GenBank under the Accession Number (AN) VAUR00000000 by the pipeline available in Prokka 1.12 [34] with the dbCAN database of carbohydrate active enzymes as reference (http://csbl.bmb.uga.edu/dbCAN).
The sequences identified as GH19 were named *lys177* and *lys188* (lys stands for lysozyme and the numbers stand for the amino acid length of the predicted protein sequences).

Enzymes expression and purification

The *lys177* and *lys188* gene sequences were codon optimized for expression in *E. coli*, synthesized and cloned into pET-21a expression vector by GenScript USA Inc. (Piscataway, NJ 08854, USA). Both genes are flanked by *Nde*I and *Xho*I restriction sites, and harbour 18 supplemental nucleotides for 6xHis-Tag at their C-terminus. The expression vector was transformed in *E. coli* DH5α (EMD Millipore, Billerica, MA, USA) for amplification and then transferred into *E. coli* BL21[DE3] cells (EMD, Millipore, Billerica, MA, USA) for heterologous production. Transformants were grown overnight at 37°C in 2 mL Lysogeny Broth (10 g/L tryptone, 5 g/L yeast extract, 5 g/L NaCl) and then diluted 1:25 in 50 mL of Zym-5052 medium [234] and incubated overnight at 20°C. Media contained ampicillin 100 mg/L.
Recombinant His-tagged LYS177 and LYS188 were extracted as described in [235], and purified by immobilized-metal affinity chromatography (IMAC) on Ni/NTA agarose resin (Jena Bioscience, Jena, Germany) at 4°C after two washing steps at 10 and 20 mM imidazole and elution in 250 mM imidazole, pH 8.0 (complete protocol of purification in **Tab. S1**).

Protein concentration was determined by the protein Bradford assay (Bio-Rad, California, USA), using bovine serum albumin as a standard. Samples containing highest protein concentrations were buffer exchanged twice by gel filtration on PD10 column (GE Healthcare, Little Chalfont, UK) against 80 mM potassium phosphate buffer, pH 6.5.
Whole cell extracts, soluble and insoluble protein fractions, and IMAC purified fractions were loaded on 14% acrylamide Tris-Glycine SDS/PAGE with BLUeye Prestained Protein Ladder by GeneDirex Inc. as the standard. After electrophoresis, gels were stained with Coomassie dye (Bio-Rad).

Enzymes characterization

a) Lysozyme activity assay

Lysozyme activity was measured spectrophotometrically in Euroclone Primo® Multiwell plates 96 (Pero, Italy), by a VICTOR Multilabel Plate Reader (PerkinElmer, Waltham, Massachusetts, U.S). ≈ 10 mg/mL cells of the Gram-positive bacterium *Micrococcus lysodeikticus* (Merck KGaA, Darmstadt, Germany) were suspended in 270 µL of 80 mM potassium phosphate buffer, pH 6.5, so that in a single well A_{600} was between 0.6 and 1. Thirty microlitersµL (≈ 2 µg) of enzyme solution were added to the reaction mix and mixed by pipetting. A_{600} was recorded at 10 s intervals up to 10 min. The slope of the linear regression between A_{600} and time (min) was used for the calculation of ΔA_{600}/min. Reactions were carried out at 30°C and the PD10 buffer was used as blank. In this assay, activity (U) is defined as the amount of enzyme that induces a decrease of 0.001 A_{600} per min, due to *Micrococcus* cell lysis following wall degradation elicited by the enzyme. Specific activity is defined here as U/mg normalized for the reaction volume, according to the formula presented in [236]. HEWL (Hen Egg White Lysozyme, Merck KGaA) was the positive control. Activity was recorded in the 3-9 pH range and in the 5°C- 65°C (pH 6.5) temperature range. Measures were taken in biological and technical triplicates.

Lysozyme activity of LYS177 towards Gram-negative bacteria was tested on *E. coli* BL21 cells treated with EDTA to increase the outer membrane (OM) permeability as described in [113]. The reaction was started by adding 30 µL LYS177 (≈ 400 ng) to 270 µL of EDTA-treated *E. coli* BL21 suspended in 80 mM K-phosphate pH 6.5 suspension. Incubation was performed at 30°C in Euroclone Primo® Multiwell plates 96. Specific activity was calculated in biological triplicates as above. HEWL was used to compare the performance in the same assay conditions.

b) Chitinolytic activity assay

Chitinolytic activity was measured on the low molecular weight chromogenic substrates 4-nitrophenyl N,N′-diacetyl-β-D-chitobioside (pNP-chitobioside) and chitin azure (CA) by Merck KGaA.
0.9 mM pNP-chitobioside was dissolved in 270 µL 100 mM sodium acetate, pH 4.8 and 6.0, or 100 mM potassium phosphate buffer, pH 7.0 or pH 8.0. Reactions started when 30 µL of enzyme solution (containing either 6 µg or 60 µg enzyme) was added to the reaction mixture. After 1h at 25°C 600 µl of sodium carbonate 0.37 M was added to stop the reaction. A_{420} was used to calculate the activity based on the molar extinction coefficient of released 4-Nitrophenol at pH 10.
3 mg/mL CA 270 µl suspension in 100 mM sodium acetate buffer (pH 4.6) or 100 mM sodium phosphate buffer (pH 7 and 8) was added with 30 µl of enzyme solution and stirred at 25°C for 1h

before heat inactivation [237]. The reaction mixture was centrifuged at 13,000 g for 5 min and the absorbance of the supernatant was measured.
Each measure was in triplicate. The standard chitinolytic cocktail from *Streptomyces griseus* (Merck KGaA) was used at different concentrations as the positive control.

c) Circular dichroism (CD) spectroscopy

Measurements of protein samples (6 µM) in 80 mM potassium phosphate buffer, pH 6.5 were performed in biological triplicates at 4°C by a spectropolarimeter J-815 (JASCO Corporation, Easton, USA) in a 1-mm path-length cuvette at variable wavelength in the far-UV range (195 - 260 nm). Other parameters were: scanning speed 20 nm/min, bandwidth 1 nm, digital integration time per data 2 s and data pitch 0.2 nm. All spectra were corrected for buffer contribution, smoothed twice by the Means-Movement algorithm and averaged among different biological replicates.
Thermal denaturation spectra were obtained by measuring the CD signal at 222 nm fixed wavelength when progressively heating the sample from 5°C to 70 °C. Measurements were performed with a data pitch of 0.2 °C and a temperature slope of 5 °C/min.
Molar mean ellipticity per residue was calculated according to the formula:

$$[\theta] = \frac{3300m \cdot \Delta A}{c \cdot n \cdot l}$$

where ΔA is the difference in the absorption between circularly polarized right and left light of the protein corrected for blank, m is the protein molecular mass in Daltons, l is the path length (0.1 cm), c is the protein concentration in mg/mL and n is the number of residues [231]. Absolute CD signals were converted to percentage (%) with respect to maximum and minimum values, and the scatterplot with temperature was fitted with a Boltzmann distribution to estimate the thermal denaturation midpoint (T_m).

d) Thermal stability

Relative lysozyme specific activity and relative CD signal at 222 nm were determined after incubation of LYS177 at 4°C, 20°C and 37°C at the pH optimum (pH_{opt}) in 80mM potassium phosphate buffer. Measures were recorded at 25°C after 4h, 8h and 24h in the first day of incubation and then after day 2 and 4, and every 4 days up to 16 days. Results were averaged over three biological replicates.

In silico analysis

a) Sequence analysis

The amino acid sequences of LYS177 and LYS188 were Blast - searched in the UniProtKB/Swiss-Prot database to obtain the closest experimentally characterized protein. SignalP 5 (http://www.cbs.dtu.dk/services/SignalP/, [238]) was used for detecting the presence of signal peptides. The sequenced genome of the isolated strain was scan-searched with Phaster [239], in order to detect if the coding genes were located within prophagic regions.

b) Phylogenetic analysis

Multiple alignments and phylogeny analysis were performed with a group of biochemically characterized GH19 selected from CAZy (http://www.cazy.org) and UniProt databases.

A starting approximate alignment was built with the E-ins-I algorithm of Mafft 7.313 [240]. All the accessory domains (not containing the GH19 catalytic domain) were then manually trimmed. A Bio-Neighbour Joining [241] starting tree was generated from this alignment through *Phylogeny.fr* web service (http://www.phylogeny.fr/one_task.cgi?task_type=bionj). These results were refined in a Bayesian analysis by Bali-Phy 3.4 [242]. 6 independent Monte Carlo Markov chain analyses were run and stopped after 40000 cycles, when sampled parameters got to convergence and good mixing according to the manual guidelines (http://www.bali-phy.org/README.html#mixing_and_convergence). In order to eliminate the background noise at the beginning of the run, the first 50% of samples was discarded. Each analysis was performed at default parameters priors with an LG empirical substitution rate matrix [243] and a rs07 [244] insertion/deletion model. The resulting unrooted tree is the majority consensus from all the samples collected during runs.

The position of the root was inferred with a parsimony-based approach, minimizing the costs of duplication, transfer, loss events under a defined species phylogeny by RANGER 2 [245], using the previously obtained unrooted phylogeny and a chronogram tree of the species in which each branch represents the evolutionary time. This tree was generated through the "Time Tree of Life" website (http://www.timetree.org/). Three different cost combinations for duplication (D), transfer (T) and loss (L) ([D-T-L]: [2-3-1], [3-3-1] and [2-4-1]) were used to select the optimal position of the root. Each analysis was repeated 100 times. The position of the root was considered reliable if optimal (minimum number of costs) in all attempts.

c) Conservation analysis and modelling

The multiple alignment from the previous step including LYS177, LYS188 and other GH19 endolysins was visualized with SeaView version 4.7 [246]. Conserved residues important for substrate binding in other GH19 [111, 114, 139, 213] were annotated.

Rate4Site (VERSION 2.01) detects conserved amino-acid sites by computing the relative evolutionary rate for each site with respect to a multiple sequence alignment [247]. Consurf 3.1 [59], http://consurf.tau.ac.il/2016/) uses Rate4Site relative rate scores and splits them in 9 bins from the "fastest" (assigned to 1) to the "slowest" (assigned to 9). Results are plotted on an alignment and on a reference structure. The above mentioned multiple alignment was used as input in the web server implementation of Consurf [248] with a LG substitution rate matrix [243] and an empirical Bayesian approach. If half or less sequences in the alignment contained gaps for a specific site, the "fastest" evolving score was assigned to the corresponding alignment column. *Salmonella Typhimurium*-infecting phage SPN1S endolysin [141], the only sequence with a structure available in protein data bank (PDB code: 4OK7) was taken as the reference structure.

The same structure was also used for building a model of LYS177 (≈ 39% identity) by RaptorX [197], one of the best performing homology modelling servers (available at http://raptorx.uchicago.edu) and with the most complete benchmark available in the CAMEO community project (https://www.cameo3d.org/sp/1-year/?to_date=2018-12-08).

The most accurate model was energy minimized with CHARM22 forcefield [249] in gromacs 2018.4 [250], then visualized and structurally aligned to the reference template with pymol 2.1 (http://www.pymol.org/), in order to predict the structural differences with respect to LYS177.

Deposited sequences

The 16S rDNA sequence of the isolated bacterial strain is deposited in the GenBank database (http://www.ncbi.nlm.nih.gov/genbank/) under the AN MH177769. The gene sequences of wild type lys177 and lys188 are deposited in the GenBank database under the AN MK926465 and MK926464, respectively. Codon optimized sequences are deposited under AN MN243085 (LYS177) and MN243086 (lys 188).

Results

Molecular identification of the bacterial isolate and of glycoside hydrolase coding sequences

The *Pseudomonas sp.* Ef1 bacterial isolate was compared to other bacterial species based on its 16S rDNA sequence. Highest identity values were found with the 16S rDNAs from *Pseudomonas koreensis* Ps9-14^T(99.77%, [251], *P. reinekei* (99.53%) and *P. granadensis* (99.50%).

Scanning the entire *Pseudomonas* Ef1 annotated genome (Ramasamy et al**.,** Microbial Resource Announcement, submitted) we identified two GH19 sequences that we named *lys177* (531 bp) and *lys188* (564 bp). *Lys177* is 98% identical to the homologous sequence from *P. koreensis*, while *lys188* is 92% identical to the sequences from *P. moraviensis* and *P. putida*, and 91% identical to the glycoside hydrolase from *P. koreensis*.

Predicted amino acid sequences of LYS177 and LYS188 share 65% sequence identity to each other. Their closest match in the UniProtKB/Swiss-Prot database is the GH19 Endolysin A from *Mycobacterium* phage D29 (AN O64203). The catalytic domain of Endolysin A is 37% and 36,8% identical to LYS177 and LYS188, respectively. It is interesting to note that both *Lys177* and *lys188* genes are inserted in a complete prophagic region of the host *Pseudomonas* Ef1, according to Phaster. SignalP 5 did not predict the presence of any signal peptide.

LYS177 and LYS188 are cold-active, thermolabile glycosidases with lysozyme activity

The yield of recombinant His-tagged LYS177 and LYS188 (**Fig. S1**), determined after affinity chromatography purification and two buffer exchange steps, was respectively ≈ 2.5 mg and ≈ 1 mg per 100 mL of culture.

Since the GH19 family groups proteins with either chitinase or lysozyme activity, we assayed recombinant proteins for both using the protocols described in the "Materials and Methods" section. While the two enzymes did not show any activity on neither of the two chitinase-specific substrates, both were active as lysozymes. In this specific assay, LYS177 pH_{opt} was 6.5 (**Fig. 1a**) and its temperature optimum (T_{opt}) 35°C (**Fig. 1b**). At 5°C LYS177 retained 40% activity, a hallmark of cold activity [183].

Under optimal reaction conditions, the specific activity of LYS177 was 1877 ± 99 U/mg. In the same assay, the activity of LYS188 was too low to be reliably calculated. For this reason the concentration of the buffer was lowered to 50 mM in order to increase the assay sensitivity (Levashov et al., 2010). pH_{opt} and T_{opt}, as well as residual activity at low temperature, did match those displayed by LYS177. Nevertheless, the temperature dependent loss of activity was sharper, with only 30% residual activity at 35°C (**Fig. 1c-d**). Specific activity under optimal conditions was 142 ± 20 U/mg.

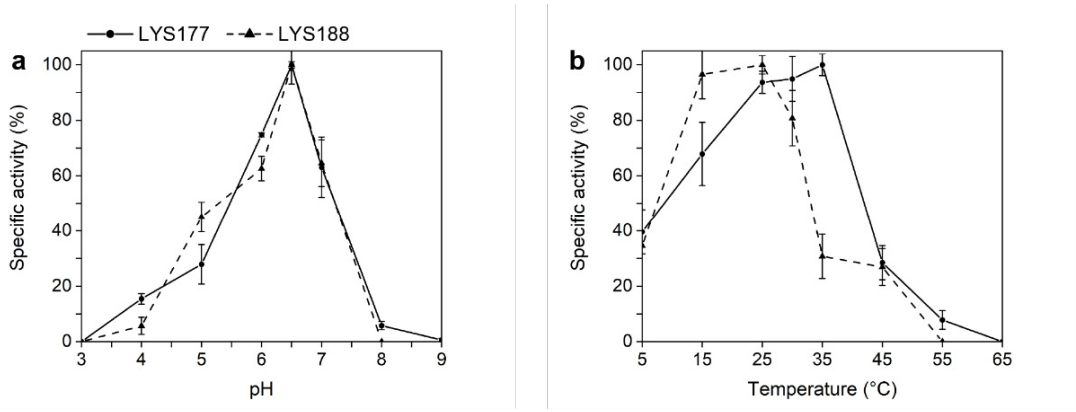


Fig. 1 Enzyme activity. The effects of pH (**a**) and of temperature (**b**) on the specific activity of LYS177 and of LYS188 was detected by means of a turbidimetric lysis assay with *M. lisodeikticus* cells. Error bars indicate standard deviations of three independent biological replicates.

As both proteins resulted to be temperature sensitive, their secondary structure was investigated by circular dichroism performed at 5°C (**Fig. 2**). CD spectra displayed a bimodal shape, with two local minima at ca. 208 and 222 nm. The observation that the minimum peak at 222 nm was slightly more negative than that at 208 nm, suggested (according to [252] an abundance of α-helices and strong inter-helix interactions. These data are consistent with a mainly alpha compact globular domain. The secondary structure content of both enzymes progressively decreased with temperature. Above 45°C, the CD signal dropped abruptly because of protein aggregation, which was detectable also as macroscopic precipitates in the cuvette. The T_m, determined at fixed 222 nm wavelength, was ca. 52°C for LYS177 and ca. 45°C for LYS188 (**Fig. 2b** and **d**). To investigate thermal stability over time, LYS177 residual activity and residual secondary structure were measured after incubation at 4°C, 20°C and 37°C, covering the conditions usually applied to study the stability of psychrophilic enzymes, and increasing time of exposure. The activity and secondary structure content of samples incubated at 4°C were stable over several days (**Fig. 3a** and **b**) after a drop of 35% of activity recorded in the first hours of the experiment (**Fig. 3a**). By contrast, both parameters decreased with time upon incubation at 20°C and 37°C, and half-lives of about 8 days and 2 days, respectively, were recorded.

Furthermore, we performed preliminary assays of LYS177 for lysozyme activity on both Gram- positive (*Micrococcus lysodeikticus)* and Gram-negative (permeabilized *E. coli* BL21) bacteria in comparison with the commercial HEWL. As shown in **Fig. 4** while LYS177 was by far less effective than HEWL (1877 ± 99 U/mg for LYS177 and 28981 ± 925 U/mg for HEWL) on *Micrococcus* cells, it performed higher specific activity on *E. coli* at pH 6.5 and 30°C (LYS177 436 ± 29 U/mg, HEWL 358 ± 101 U/mg). This behavior may ground in the Gram-negative background from which this enzyme origins.

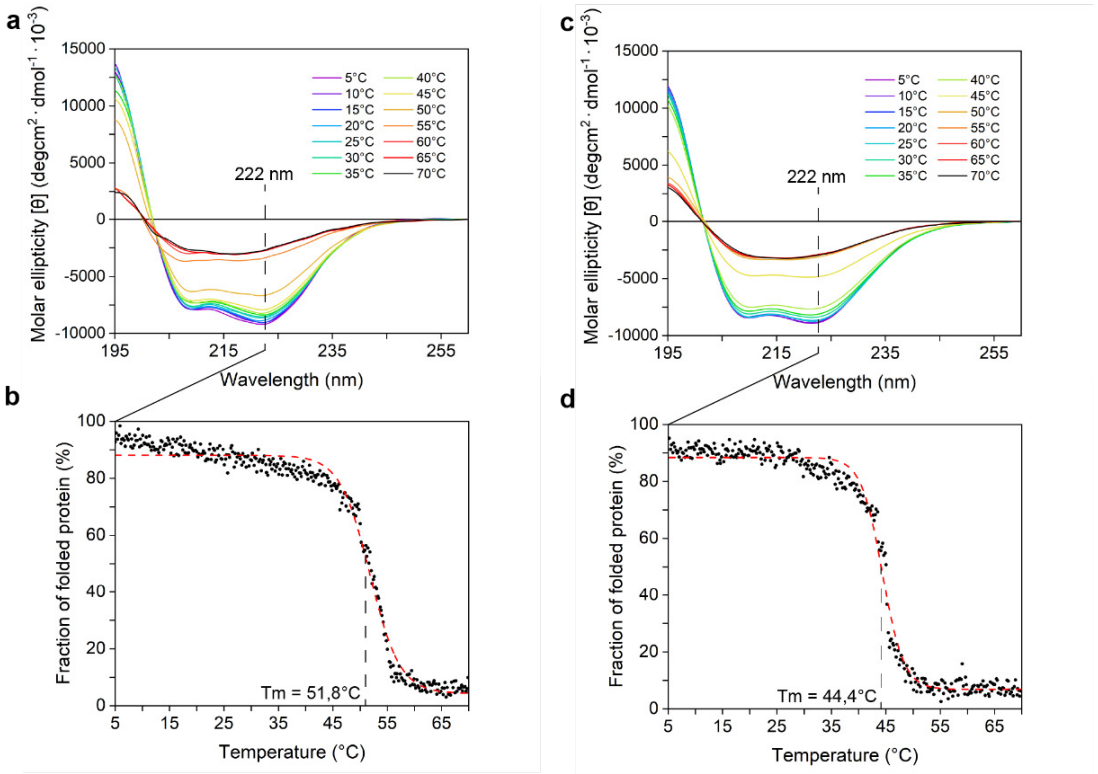


Fig. 2 CD spectroscopy analysis. Far-UV CD spectra (**a** and **c**) of LYS177 and LYS188 recorded at different temperatures. Thermal unfolding (**b** and **d**) of LYS177 and LYS188. Ellipticity values were collected at fixed 222 nm wavelength during heating from 5°C to 70°C. Initial CD signal was taken as 100% for normalization. The Boltzmann fitting was used to estimate Tm. Data are the average of three independent experiments.

LYS177 and LYS188 belong to a specific group of phage and prophage endolysins

We studied the evolution of the two novel glycosydases within the GH19 family by Bayesian phylogenetic analysis limited to GH19 sequences present in the CAZy and UniProt databases and for which biochemical characterization is available. We identified five major monophyletic clusters consistent with specific functional and taxonomic groups (**Fig. 5**). Four out of them include bacterial chitinases, Proteobacteria chitinases, and two clusters of plant chitinases differing in the number of substrate binding loops (the terms “loopful” and “loopless” refer to [207]). The chitinase from a Moss [111] clusters at the basis of plants GH19, whereas that from a fungus [223] shares a common ancestor with a derived group containing phage/prophage endolysins with lysozyme activity and ORF69 from cyanophage Ma-LMM01, whose chitinolytic activity was reported [137]. LYS177 and LYS188 nest in this group confirming the classification of the two enzymes as endolysins.

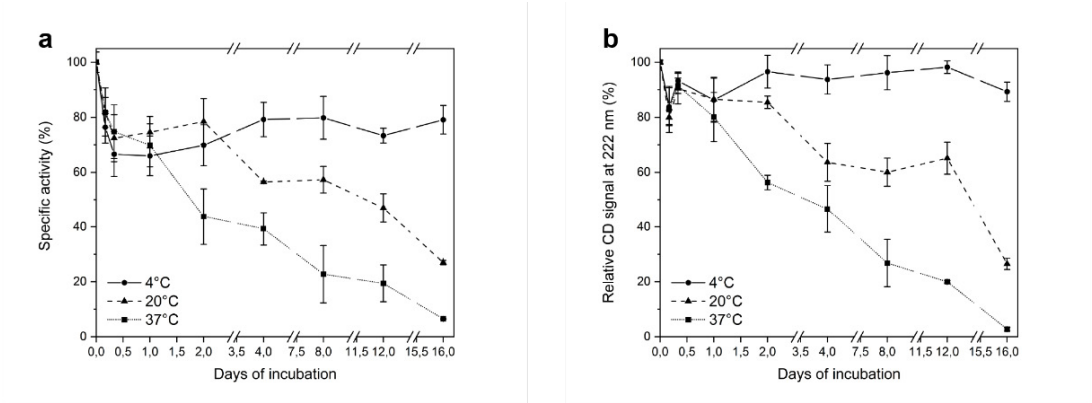


Fig. 3 Thermal Stability of LYS177. Relative specific activity (**a**) and relative CD signal at 222nm (**b**), interpreted as the relative content in secondary structure, after incubation of the enzyme at different temperatures.

Fig. 6a reports the multiple alignment of GH19 endolysins annotated by Consurf analysis (per site relative evolutionary rates, described in more detail in **Tab. S2**) integrated with information about the position of the active site and of the substrate binding regions in other known GH19 proteins. To better clarify the picture, we considered the alignment positions with respect to the only GH19 endolysin structure available (**Fig. 6b**), the bacteriophage SPN1S endolysin [141]. SPN1S endolysin consists of two domains, a large "lobe" and a smaller one containing a three-helix bundle hosting the peptidoglycan-binding domain. In this structure the catalytic dyad (GLU-49 and GLU-58) responsible for the inversion catalytic mechanism of GH19 is strictly conserved and faces the groove between the two lobes. Most active site residues (corresponding to positions of HIS-50, THR-97, ASN-101, VAL-155 and ASN-156, with the exception of GLY-53) relevant for the interaction with chito-oligomer substrates [111, 139] are also conserved, and belong to Consurf categories 8 or 9, the maximal score for residue conservation, suggesting their slow evolutionary rate. As these residues also point towards the active site (**Fig. 6b**), this observation surmises they may have retained the same substrate binding function experimentally demonstrated in chitinolytic enzymes, considering that chito-oligomers are not too different from the sugar chain component of peptidoglycans.

It is interesting to note that in the overall alignment several indels are interspersed with short hotspots of highly conserved regions. These "slowest" evolving sites are mainly located within and around the catalytic cleft (**Fig. 6b**), while most residues on the surface of the two "lobes" were assigned to faster relative rate categories. In particular, this is true for the peptidoglycan binding 3-helix bundle described in [141], not conserved in the alignment. Worth of notice is that the energy minimized 3D model of LYS177 (**Fig. 6c**) is in good agreement with the structure of the template, but it lacks two out of the three helices of this motif, and the one that is structurally overlapped is completely different in primary sequence.

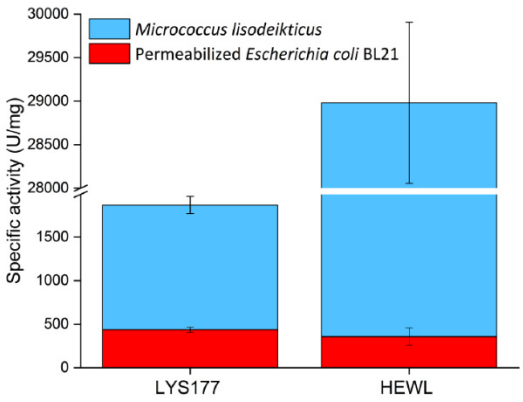


Fig. 4 Activity against Gram-positive and Gram-negative bacteria. The bar plot displays the absolute values of the specific activity of LYS177 and HEWL against *M. lisodeikticus* and OM permeabilized *E. coli* BL21 cells.

Discussion

The two glycoside hydrolases described in this paper are endowed with lysozyme activity and are classified as endolysins based on the evolutionary analysis performed and of the antimicrobial activity they display. Because of their Antarctic origin, we were interested in defining temperature dependence and temperature stability. A first remark was that the temperature supporting highest activity (35°C for LYS177and 25°C for LYS188) though moderate is higher than the temperature of growth of the bacterium that should approach 0°C. The same peculiarity has been reported for several enzymes form psychrophiles [253, 254]. Both endolysins are consistent with the canonical definition of cold-active enzymes since they maintain high specific activity at low temperature and are thermolabile. By the experiments of incubation at increasing temperature and time, we observed that the loss of activity parallels the reduction of protein secondary structure. This behaviour, shared by several cold-adapted enzymes, suggests that conformational flexibility spreads over the whole protein structure. In other cases, loss of activity precedes denaturation, in agreement with a highly flexible active site embedded in an overall stable structure. The issue of thermolability of cold-active enzymes is far to be fully unveiled. While some enzymes lose their activity at mild temperature within a few hours only [194], in other cases robustness to heat is comparable to that of mesophilic or even thermophilic enzymes [195]. Such cases seem to question the paradigm “localized flexibility allows for cold-activity”. However, cold activity might depend on increased flexibility restricted to very small protein regions or loops that are not enough mobile to unfold separately from the protein structure and are therefore difficult to detect [166].

Based on the GH19 tree topology, we can hypothesize that the common ancestor of GH19 proteins was most probably a chitinolytic enzyme, while lysozyme activity arose later. Our analysis does not allow inferring which type of organism harbored the common ancestor of all GH19s, because the first bifurcation from the tree root is between proteins mainly from Actinobacteria and those from a composite group including bacterial, plant, phage and fungal proteins. Nevertheless, data support the

hypothesis of horizontal gene transfer from plants to Proteobacteria, before the diversification of "loopful" plant chitinases.

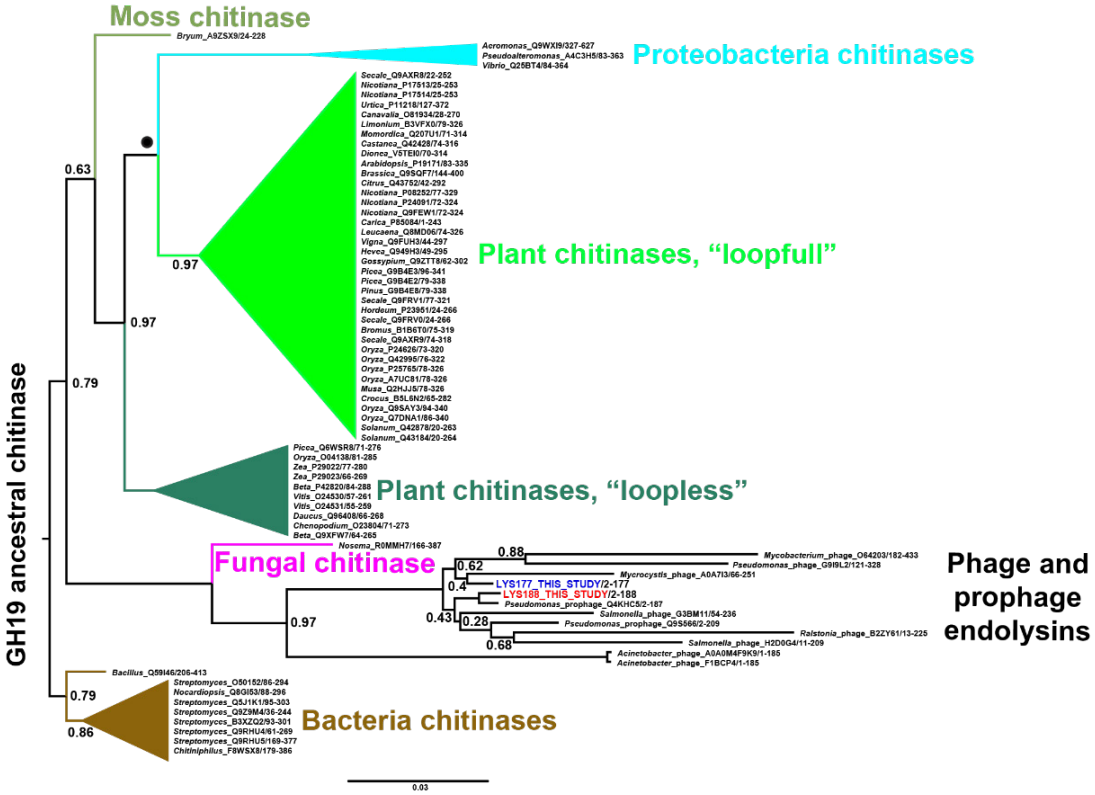


Fig. 5 Rooted phylogenetic tree of characterized GH19 proteins. Each tip name represents the organism *genus* separated by an underscore from the Uniprot AN and by a slash from the sequence start and end position. In the Endolysins cluster the two sequences considered in this study are highlighted in blue and red. Other clusters/tips reported in "Results" section are coded with different colours and their internal relationships are collapsed for visualization purpose. The "●" symbol indicates a hypothetical horizontal gene transfer. Decimal numbers at internal nodes indicate posterior probabilities only if lower than 1. The branch lengths are proportional to the expected number of substitutions per site.

The evolutionary scenario of the GH19 family presented in this work is different from the most recent family revision [114], as it does not surmise a plant origin and secondary transfer to bacteria. We are aware that, in the lack of a dated tree and a comprehensive sequence sampling, relevant elements are still missing for confirming the inferred rooting and horizontal gene transfer events. Nevertheless, the big average pairwise distance between all the identified groups (≈ 60,1 ± 7,3%) indicates that GH19 sequences have a very old history. The average pairwise distance between endolysins is also very high (≈ 59,4 ± 14,7%), suggesting that functional specialization appeared very early (or that molecular evolution was very fast). The number of GH19 endolysins described to date is too poor and the alignment variability too high to allow drawing functional hypotheses on specific parts of LYS177 or LYS188 sequences, which also lack two out of three helices of the peptidoglycan binding motif. However, the position of the variable amino acids mapped on the reference structure suggests that these residues are not essential for catalysis and active site substrate positioning.

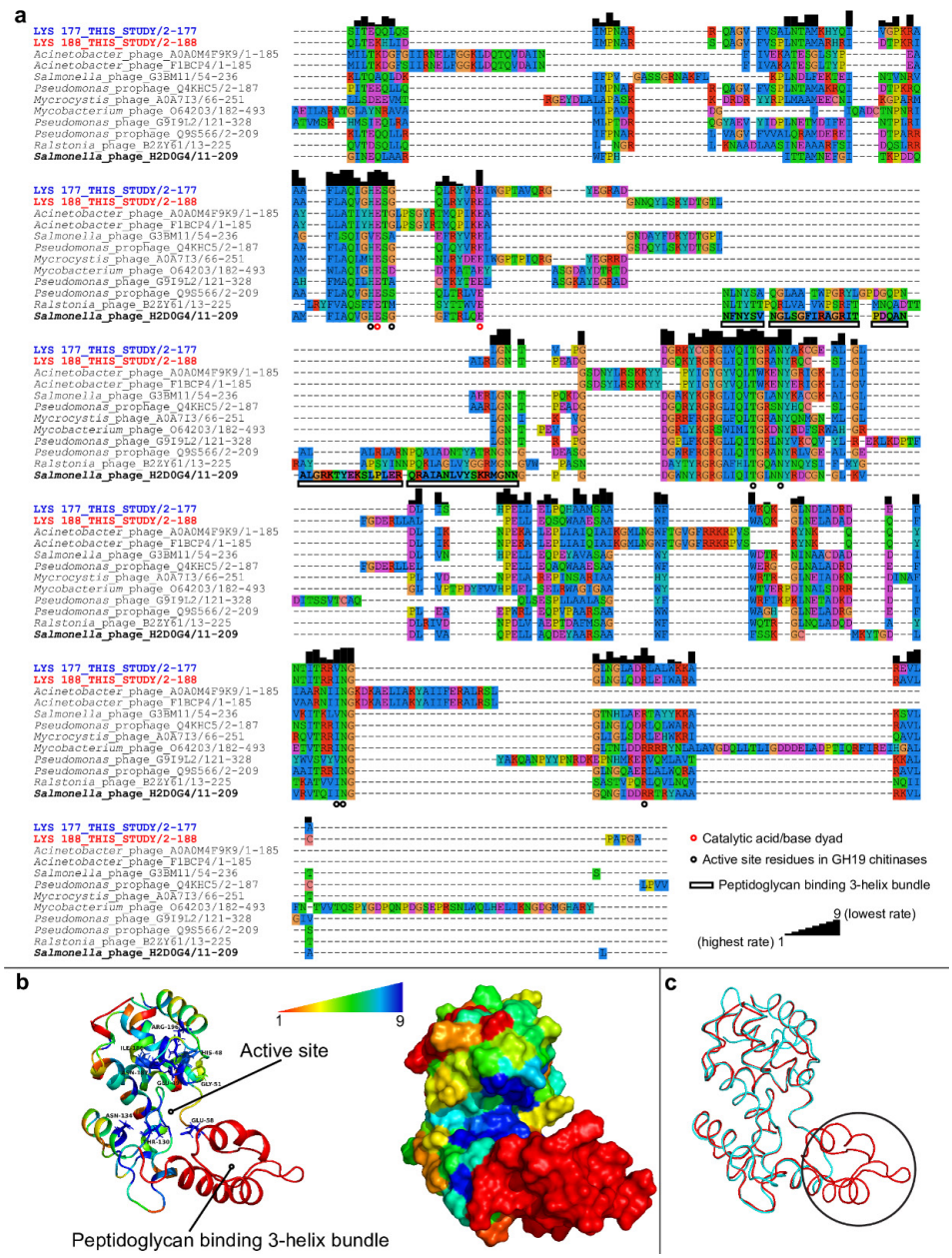


Fig. 6 Endolysin cluster alignment and structural annotation. The multiple alignment (**a**) of sequences belonging to the GH19 endolysin cluster (see **Fig. 5**) was annotated based on Consurf analysis and literature information. Evolutionary rates per site are visualized in black on the top of the alignment, divided between the 9 different Consurf categories. Substrate binding residues conserved in GH19 and the peptidoglycan binding 3-helix bundle of *Salmonella*_phage_H2D0G4 endolysin (the reference structure) are marked. The colour code for residues in the alignment follows their biochemical properties: blue = non-polar; red = basic; green = polar; violet = acid; yellow = proline; orange = glycine. Annotations were plotted onto the reference structure represented in cartoon and surface style (**b**): the substrate binding conserved residues are visualized in sticks on the left image while the surface is showed on the right. The colour code flows the legend gradient from the most conserved (slowest rate) in blue, to the less conserved (fastest rate) in red. In (**c**) the backbones of the reference structure (red) and the predicted model of LYS177 (cyan) are structurally aligned (RMSD = 0.088 nm). A circle highlights the major structural difference.

Instead, they could participate in the plasticity necessary for co-evolutionary phage-host interactions [196]. In this scenario, more work is required to understand if such surface residues could have played a role in host-specific successful invasion. This information might be relevant also for the design of specific antibacterial agents. Moreover, the observed activity in the degradation of both Gram-positive and Gram-negative bacteria, contributes to the emerging strategy of exploiting “enzybiotics” [255] in the control of microbial growth at low temperature.

Acknowledgements

M.O. acknowledges a PhD fellowhip by the University of Milano-Bicocca.

Compliance with Ethical Standards

Conflict of Interest:

M. Orlando declares that he has no conflict of interest.
S. Pucciarelli declares that she has no conflict of interest.
M. Lotti declares that she has no conflict of interest.

Funding

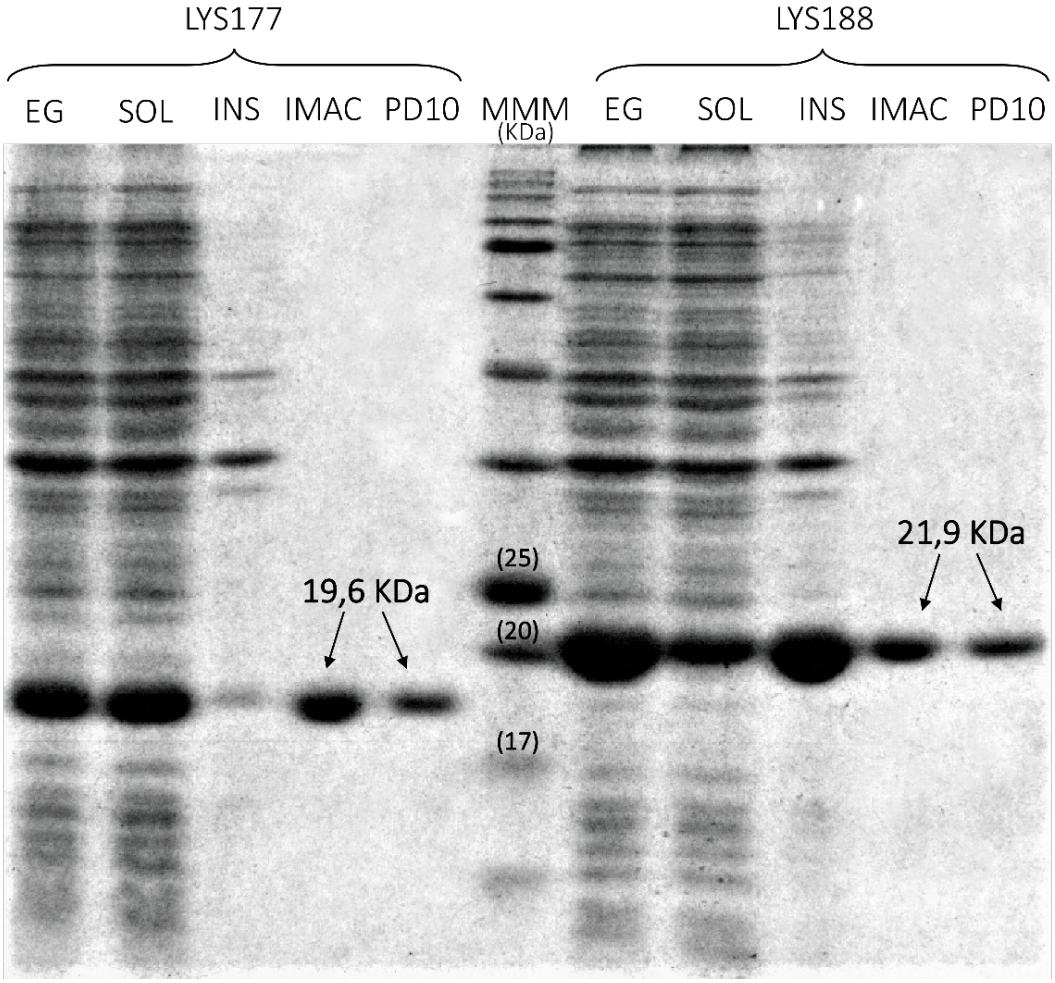
This work was supported by the Project EU H2020-MSCA-RISE MetABLe-645693 co-ordinated by S.P. M.L. acknowledges support by FA (Fondo di Ateneo) of the University of Milano-Bicocca (grants 2015-ATE-0060, 2016-ATE-0134).

Ethical Approval:

This article does not contain any studies with human participants or animals performed by any of the authors.

Supplementary Material

Fig. S1 SDS/PAGE of recombinant proteins purified by affinity chromatography (IMAC) and PD10 exchanged. MMM: molecular mass marker. For both samples, the first three lanes from the left contain the total (TOT), soluble (SOL) and insoluble (INS) fractions from the lysates after production in Zym-5052 medium (see "Materials and Methods"). IMAC and PD10 lanes contain ≈ 1,5 µg of protein. The sequence-based predicted molecular weight of both proteins is pointed by black arrows on each corresponding IMAC and PD10 lane.



Tab. S1 IMAC and PD10 gel filtration purification protocol.

1. 5 min centrifugation at 6000 rpm of 50 ml *E. coli* BL21 cell colture (O.D. ≈ 7)		
Step	**Buffer**	**Volume (µL)**
Cell resuspension	Lysis buffer (10mM imidazole, pH 8 sodium phosphate buffer 50 mM, NaCl 300 mM)	4000
2. Resuspended cells are lysed and the soluble content is collected from the supernatant after centrifugation for 15 min at 8000 rpm. The soluble fraction is loaded on an IMAC column		
Step	**Buffer**	**Volume (µL)**
1st Flowthrough	Lysis buffer	4000
2nd Flowthrough	Wash buffer (20mM imidazole, pH 8 sodium phosphate buffer 50 mM, NaCl 300 mM)	4000
1st Eluition	Eluition buffer (250mM imidazole, pH 8 sodium phosphate buffer 50 mM, NaCl 300 mM)	500
2nd Eluition	"	500
3rd Eluition	"	500
4th Eluition	"	500
5th Eluition	"	500
6th Eluition	"	500
7th Eluition	"	500
3. Collection of IMAC eluted fractions with the target enzyme, after measurement of protein concentration through Bradford assay. Loading of a 1st PD10 gel filtration column		
Step	**Buffer**	**Volume (µL)**
1st Eluition	Final buffer (pH 6.5 80 mM potassium phosphate buffer)	750
2nd Eluition	"	750
3rd Eluition	"	750
4th Eluition	"	750
5th Eluition	"	750
6th Eluition	"	750
7th Eluition	"	750
8th Eluition	"	750
9th Eluition	"	750
4. Collection of gel filtration eluted fractions with the target enzyme, after measurement of protein concentration through Bradford assay. Loading of a 2nd PD10 gel filtration column		
Step	**Buffer**	**Volume (µL)**
1st Eluition	"	750
2nd Eluition	"	750
3rd Eluition	"	750
4th Eluition	"	750
5th Eluition	"	750
6th Eluition	"	750
7th Eluition	"	750
8th Eluition	"	750
9th Eluition	"	750
5. Final collection of PD10 gel filtration eluted fractions with the desired product, after measurement of protein concentration through Bradford assay		

Tab. S2 Per residue conservation scores. The table provides the Consurf conservation scores at each position of the reference Salmonella_phage_H2D0G4 endolysin sequence, plotted on the alignment and on the structure in **Fig. 6a** and **b**. The scores range from "1", indicating the most variable (fastest rate category), to "9", indicating the most conserved (slowest rate category) sites.

Residue n°	Residue type	Consurf Score	Positions in the alignment	Residue variety
1	G	1	12/12	S,P,G,L,H,M,Q,K
2	I	4	12/12	M,V,L,I
3	N	6	12/12	A,N,T,L,S
4	E	3	12/12	D,T,I,E,Y,Q
5	Q	2	12/12	Q,K,N,A,S,E
6	L	6	12/12	Q,R,H,L,D,E
7	A	5	12/12	G,A,V,L
8	A	1	12/12	F,M,R,Q,A,I,L,V,D
9	R	1	12/12	A,R,K,Q,D,T,G,S
10	W	5	10/12	W,M,L,I
11	F	3	9/12	A,M,F,L
12	P	8	9/12	P
13	H	4	10/12	A,N,H,V,T
14	I	6	11/12	A,I,L
15	T	6	11/12	M,Q,N,T
16	T	3	11/12	D,T,E,A,R
17	A	7	11/12	T,L,S,A
18	M	7	12/12	L,F,M,G,A
19	N	1	11/12	K,A,N,D,L,E
20	E	1	11/12	H,R,K,E,S,I,T
21	F	1	11/12	H,C,F,R,Y,T
22	G	1	11/12	P,G,N,Q,R,S,E
23	I	9	10/12	I
24	T	4	10/12	N,K,D,V,T
25	K	3	10/12	T,N,G,Q,K
26	P	6	10/12	S,V,P
27	D	1	10/12	K,N,A,L,D
28	D	7	12/12	D,E,R
29	Q	1	12/12	I,V,R,Q,A,M
30	A	9	11/12	A
31	M	5	11/12	A,G,M,H,Y
32	F	8	12/12	L,F,W
33	I	7	12/12	M,I,V,L
34	A	9	12/12	S,A
35	Q	9	12/12	T,Q
36	V	5	12/12	I,S,V,L
37	G	3	12/12	G,M,F,L,Y
38	H	8	12/12	F,H,V
39	E	9	12/12	E
40	S	8	12/12	T,S
41	G	4	12/12	S,D,A,G,M
42	G	3	12/12	D,T,E,N,Q,S,G,C
43	F	6	12/12	L,Y,F,M
44	T	3	12/12	Q,K,R,T
45	R	2	12/12	R,A,P,T,Y
46	L	1	12/12	W,T,I,L,D,V
47	Q	1	12/12	V,E,R,Q,K,A

48	E	9	12/12	E
49	N	1	3/12	N
50	F	1	3/12	F,L
51	N	1	3/12	T,N
52	Y	1	3/12	Y
53	S	1	3/12	S,T
54	V	1	3/12	T,V,A
55	N	1	3/12	Q,N
56	G	1	3/12	G,R
57	L	1	3/12	L
58	S	1	3/12	S,V,A
59	G	1	3/12	G,A
60	F	1	1/12	F
61	I	1	3/12	I,T,V
62	R	1	3/12	R,W
63	A	1	3/12	P,A
64	G	1	3/12	S,G
65	R	1	3/12	R
66	I	1	3/12	F,I,Y
67	T	1	3/12	T,L
68	P	1	3/12	M,P,D
69	D	1	3/12	G,N,D
70	Q	1	3/12	Q
71	A	1	3/12	A,P
72	N	1	3/12	D,N
73	A	1	3/12	A
74	L	1	3/12	L,Y
75	G	1	1/12	G
76	R	1	1/12	R
77	K	1	1/12	K
78	T	1	1/12	T
79	Y	1	1/12	Y
80	E	1	1/12	E
81	K	1	1/12	K
82	S	1	3/12	A,S
83	L	1	3/12	L,P
84	P	1	3/12	S,P,R
85	L	1	3/12	Y,L
86	E	1	3/12	E,I,A
87	R	1	3/12	N,R
88	Q	1	3/12	P,Q
89	R	1	3/12	Q,R
90	A	1	3/12	K,A
91	I	1	3/12	L,I
92	A	1	3/12	A
93	N	1	3/12	D,N,G
94	L	1	3/12	N,L
95	V	1	3/12	T,V
96	Y	1	3/12	Y
97	S	1	6/12	G,A,S
98	K	1	6/12	G,A,K,E,L,T
99	R	1	6/12	R

100	M	6	10/12	N,M,L
101	G	9	10/12	G
102	N	9	10/12	N
103	N	6	1/12	N
104	G	4	10/12	T,I,G
105	P	1	9/12	K,R,P,V,D
106	G	8	12/12	G,N
107	D	9	10/12	D
108	G	8	10/12	A,G
109	W	1	10/12	Y,W,Q,R,A,P
110	N	1	12/12	L,T,N,P,R,K
111	Y	7	12/12	Y,F
112	R	5	12/12	I,R,K,C
113	G	9	12/12	G
114	R	8	12/12	R,Y
115	G	8	12/12	S,G
116	L	5	12/12	W,A,Y,L
117	I	5	12/12	F,L,V,I
118	Q	9	12/12	H,M,Q
119	I	8	12/12	I,V,L
120	T	9	12/12	T
121	G	8	12/12	G,W
122	L	5	12/12	L,R,Q,K
123	N	5	12/12	D,L,S,E,N,A
124	N	9	12/12	N
125	Y	9	12/12	Y
126	R	1	12/12	N,A,R,Q,K,H,V,E,G
127	D	2	12/12	N,A,K,R,Q,D,L
128	C	5	12/12	M,C,F,Y,V,I
129	G	5	10/12	Q,G,S
130	N	1	10/12	V,I,E,N,R,K
131	G	3	12/12	M,F,A,G,Y,L,S
132	L	7	11/12	I,L,H
133	K	4	12/12	Y,K,R,G
134	V	1	11/12	G,R,L,V,I,E
135	D	3	11/12	A,G,P,E,D
136	L	9	11/12	L
137	V	7	9/12	I,V,E
138	A	2	9/12	D,S,P,A,N,K
139	Q	7	9/12	N,Q,H,E
140	P	9	11/12	P
141	E	6	11/12	E,L,D,W
142	L	3	12/12	K,Q,R,E,L
143	L	7	12/12	A,V,L
144	A	3	12/12	L,S,E,A,R
145	Q	6	12/12	Q,S,L,E
146	D	4	12/12	L,S,D,P,A
147	E	1	12/12	E,I,T,L,V,Q,R
148	Y	1	12/12	P,Y,N,H,W,I,D,L
149	A	9	12/12	A,S
150	A	6	12/12	V,I,A,G,F
151	R	1	12/12	E,L,I,M,A,Q,R

152	S	6	12/12	A,G,I,S
153	A	9	12/12	S,A
154	A	7	12/12	I,G,A
155	W	7	12/12	W,H,Y
156	F	6	12/12	W,F,Y
157	F	7	10/12	W,F
158	S	1	10/12	S,T,D,E,Q,K,R,A
159	S	2	10/12	R,Q,G,F,T,V,S
160	K	1	10/12	H,G,K,R,E,I
161	G	4	12/12	D,K,G,N
162	C	5	12/12	L,I,Y,C
163	M	1	1/12	M
164	K	1	1/12	K
165	Y	1	1/12	Y
166	T	1	1/12	T
167	G	1	1/12	G
168	D	5	12/12	D,E,A,Q
169	L	6	12/12	F,I,L,Y
170	V	1	12/12	Y,A,N,R,E,V,I,T
171	R	1	12/12	A,R,Q,K,W,S,T
172	V	7	12/12	I,V,A
173	T	8	12/12	R,S,T
174	Q	4	12/12	Q,K,R,N,V
175	I	3	12/12	R,Y,L,V,I
176	I	8	12/12	V,I
177	N	9	12/12	N
178	G	9	12/12	G
179	G	6	10/12	S,E,G
180	Q	3	10/12	Q,P,A,T,L
181	N	7	10/12	N,S,T,I
182	G	4	10/12	T,H,N,G
183	I	5	10/12	Q,M,L,V,I
184	D	3	10/12	K,Q,P,A,S,D
185	D	5	10/12	D,E,Q
186	R	9	10/12	R
187	R	5	10/12	T,L,V,R
188	T	1	10/12	R,Q,A,T,E
189	R	1	10/12	M,H,R,Y,I,L,V
190	Y	3	10/12	Y,L,W
191	A	2	10/12	K,Q,A,N
192	A	1	10/12	L,V,Q,R,K,A
193	A	7	10/12	A,I,T,V
194	R	5	10/12	H,N,Q,R,K
195	K	2	10/12	E,S,K,Q,G,A
196	V	7	10/12	A,V,I
197	L	8	10/12	L,I
198	A	4	9/12	V,S,T,C,A

Evolutionary plasticity of glycoside hydrolase family 19

Marco Orlando[1], Patrick C. F. Buchholz[2], Marina Lotti[1], Jürgen Pleiss[2]*

[1]Department of Biotechnology and Biosciences, University of Milano Bicocca, Milano, Italy,
[2]Institute of Biochemistry and Technical Biochemistry, University of Stuttgart, Stuttgart, Germany

* Corresponding author:
Jürgen Pleiss
E-mail: Juergen.Pleiss@itb.uni-stuttgart.de
ORCID: 0000-0003-1045-8202

Keywords: sequence-structure-function relationships; biopolymer-degrading enzymes; protein evolution; glycoside hydrolases 19.

Abstract

The glycoside hydrolase family 19 (GH19) includes chitinases and endolysins, which are interesting enzymes for the control of plant fungal pests and the treatment of multi-drug resistant bacteria. 22461 GH19 sequences were collected to establish the GH19 Engineering Database (GH19ED, https://gh19ed.biocatnet.de). The sequences were assigned to two superfamilies, chitinases (8554 sequences) and endolysins (10967 sequences). The chitinase superfamily was split into 17 homologous families. The previously established plant classes I and II were merged into a single homologous family, and class IV was split into two families. Two new plant families of enzymatically inactive proteins and 12 families from prokaryotic and eukaryotic taxa were added. The endolysin superfamily is more diverse and consists of 34 homologous families.

Despite their sequence diversity, 27 residues were conserved in both chitinases and endolysins. The two families were distinguished by 6 and 4 specifically conserved residues outside the active site, which serve as signatures to predict the substrate specificity of GH19 enzymes.

The evolution of the GH19 sequence space was investigated by phylogenetic analysis. Despite the large number of homologous families, all endolysins were confined to phages and their bacterial hosts and have a similar sequence length. In contrast, chitinases are found in many eukaryotic and prokaryotic taxa, and varied in sequence length due to four loop insertions. The GH19 catalytic domain is hypothesized to have been transferred from plants to bacteria by two horizontal gene transfers, while different accessory binding modules were associated in the plant and the bacterial lineages.

Introduction

Chitinases (EC 3.2.1.14) and lysozymes (EC 3.2.1.17) belong to the class of glycoside hydrolases (GH) and catalyze hydrolysis of the glycosidic bonds of chitin and peptidoglycan polymers, respectively [95]. Chitin, the second most abundant polysaccharide in the biosphere, is an insoluble homopolymer of β-(1–4)-linked N-acetylglucosamine (GlcNAc) monomers [102]. Peptidoglycan (or murein) is a complex polymer whose polysaccharidic component is a heteropolymer of β-(1–4)-linked GlcNAc and N-acetylmuramic acid residues and is found in the cell wall of Eubacteria [105]. Both enzymes play fundamental biological roles, chitinases in the protection towards chitin containing organisms, in the degradation of chitinous organic matter into derivative nutrient sources and in autolytic morphogenetic processes in Eubacteria and Eukaryota [103, 104], and lysozymes as antimicrobial agents in animals [107].

Glycosyl hydrolases have been assigned to 165 families in the database of Carbohydrate Active Enzymes (CAZy) [79], including seven families of chitinolytic enzymes (GH3, GH18, GH19, GH20, GH23, GH48, GH84) [103] and five families of lysozymes (GH19, GH22, GH23, GH24, GH46) [106]. Although lysozymes do not share any obvious sequence conservation, they are thought to have evolved from a common ancestor because they share a hairpin and an α-helix in their catalytic core [106]. As a matter of fact, some lysozymes and chitinases are promiscuous and show a minor activity toward chitin and murein, respectively [256-259], despite the differences in the protein fold, substrate binding residues and catalytic mechanism among families [260]. A comprehensive classification based on sequence, structure and functional properties is the prerequisite for building efficient tools for the comprehension and, in perspective, modification of these enzymes. GH19 is interesting with respect to other GHs because it collects specialized enzymes that function either as endo-chitinases [95, 104, 111] or as lysozymes [112, 113]. Therefore, we focused on the properties of the sequence and structure space of this family, as it represents an ideal opportunity to investigate the relationship between and the evolution of these two types of activities, likely to be performed by the same catalytic mechanism in GH19.

Indeed, structural studies on all GH19s revealed that these enzymes have a globular mainly-alpha fold and a catalytic core spanning a deep catalytic cleft [104]. Their proposed model of hydrolysis follows a single displacement mechanism causing an inversion of the anomeric carbon (**Fig. S1**), with two glutamic acids acting as acid and as base which activates a water molecule. The nucleophilic water molecule is coordinated by a third catalytic residue, which is usually serine or threonine [110].

In early studies, GH19 enzymes were only found in plants and were grouped into three classes, class I, II, and IV [117-119]. Class II enzymes are characterized by the absence of a carbohydrate binding domain (CBM), whereas class I and IV enzymes are linked to an accessory N-terminal CBM [122]. Class IV enzymes display deletions with respect to classes I and II, resulting in a smaller number of subsites in the catalytic cleft and a different substrate binding mode [111, 121]. GH19 identified in Actinobacteria have been suggested to originate from horizontal gene transfer (HGT) from plants, and were shown to be similar in terms of deletions to and included into class IV GH19 [124, 125, 261]. However, different CBMs are linked to chitinases in Actinobateria and in class IV plant chitinases [114, 125]. Recently, an alternative classification scheme has been proposed by dividing GH19 chitinases into

"loopful" and "loopless" chitinases, based on the presence or absence of up to six loop insertions [122], in this study named 1, 2, 3, 4, 5, and C-terminal. Despite these classifications, GH19 were also detected and characterized as chitinases in Proteobacteria [127-130] and endolysins with lysozyme activity in phages [112, 113, 134-136]. Endolysin is a generic term used to indicate many different enzymes naturally produced by bacteriophages at the end of their replication cycle to degrade the peptidoglycan of the bacterial host from within, resulting in cell lysis and release of progeny virions [14].

The main biological activity of GH19 enzymes in plants is associated with improved resistance against fungal pests [142-144, 147], especially when they contain accessory CBMs [262, 263], and against phytopathogenic bacteria [204, 257, 264], demonstrated by both *in vitro* activity assays and *in vivo* expression detection during plant immune responses. Moreover, pest tolerance was demonstrated to increase in transgenic plants in which heterologous GH19 genes were introduced [265-267]. Members of the GH19 family can also be involved in the stress response caused by wounding, drought, or high temperature [210, 268] and in the regulation of lignin accumulation during plant growth [209, 269]. As GH19s are endo-chitinases, they can be applied in the degradation of chitin to chitooligomers, active as probiotics, anti-inflammatory drugs [149, 150], and for recycling chitin extracted from shellfish biomass waste [103, 104]. Moreover, improving the performance or modifying the properties of GH19 chitinases by site-directed mutagenesis was attempted [155, 207, 270, 271]. Recently, a GH19 endolysin was shown to induce outer-membrane permeabilization on treated Gram-negative bacteria strains isolated from hospitalized patients [134], proving potential in the search of drugs to fight multi-drug resistant bacteria [156].

Thus, GH19 enzymes are interesting not only for their substrate specificity, but also for the discovery of novel enzymes with valuable applications. However, only a tiny fraction of all sequences has been experimentally characterized, yet, and functional predictions on others is mostly based on automatic annotations transferred by sequence similarity [89]. Currently, in CAZy [79] some families have been manually split into subfamilies, associated to a different substrate specificity, but for most of them, including GH19, this information is not provided [79]. Recently, the selection of interesting carbohydrate active enzymes, predicted by sequence space mining, was validated as a good strategy, after measuring activity towards a broad range of carbohydrate substrates [90]. Therefore, in order to analyze and provide a comprehensive knowledge on GH19 sequence space, we created the GH19ED, as a family database on GH19 enzymes, and assessed the class- and loop-based classification systems. Moreover, we introduced new standard numbering schemes for the GH19 domain, enabling a comprehensive comparison of per-site calculated conservation scores, and investigated the evolution of the GH19 family in relation to annotated features.

Results

1. GH19ED database setup and classification

To build the GH19ED database, 23853 BLAST hits were retrieved, by using 75 biochemically characterized GH19s, manually screened from literature, as seed sequences (**Tab. S1**), and then filtered

with the GH19 profile hidden Markov model (HMM) from Pfam, resulting in 22461 sequence and 16120 protein entries in the final version (see *Experimental Procedures* section for precise definitions of ontologies in the database).

Analysis of the length distribution of sequences containing at least one GH19 domain (**Fig. S2**) evidenced that nearly 9000 sequences contain around 200 residues and more than 1500 sequences are approx. 580 residues long. Most of other sequences are up to 1100 residues in length. A few highly modular sequences up to nearly 6000 residues were also detected. The sequences of GH19 domains were annotated, extracted from the database and clustered with a fast heuristic method in order to obtain domain centroids that shared no less than 90% identity with other members of the same cluster.

Domain-based sequence networks were built by global pairwise sequence alignment of centroids, and edges were defined between two centroids (the nodes) if they shared at least 40% sequence identity. At that threshold, there were two large networks representing 19521 sequence entries (87% of all entries in GH19ED database), including all the biochemically characterized seed sequences (**Fig. S3**).

The sequences within these two main clusters (**Fig. 1**) were assigned to two separate superfamilies, chitinases (CHITs, 8554 sequences) and endolysins from phage and prophages (ELYSs, 10967 sequences), based on the fact that all seed sequences in CHITs and ELYSs were previously characterized as chitinases and endolysins, respectively. The remaining sequences, being part of small networks (less than a few tens of nodes) without characterized seeds, were included in the database, but not further analysed.

The standard numbering scheme of the CHIT superfamily was created from a profile HMM obtained aligning the sequences of 14 chitinases with known X-ray structure and other 44 biochemically characterized chitinases (**Tab. S1**). The "loopful" chitinase from *Secale cereale* (rye seed, PDB accession 4j0l) was selected as the reference for standard numbering. For this enzyme, a complete mapping of substrate binding subsites and chitinase loops is available [139]. The standard numbering covers the catalytic domain from position 24 to 266 (the first 23 amino acids are the N-terminal signal peptide). This CHIT profile HMM was used to refine the annotation of CHIT catalytic domains that were previously extracted and clustered, in order to obtain domain centroids networks of CHITs only. Edges were defined between two nodes if two centroids had at least 60% sequence identity (**Fig. 2A**). The 18 resulting clusters allowed to split CHITs into 17 homologous families (**Tab. S2**). Two of them include "loopful" chitinases (class I and II, ID 1) and "loopless" chitinases (class IV, ID 2A-B). Clusters 2A and 2B were merged, because both contain sequences characterized as class IV "loopless" chitinases from plant (2A) or from Bryophyta (2B). Two smaller homologous families contain plant proteins characterized as non-enzymatic (IDs 3-4): a chitinase-like lectin from *Urtica dioica* [132, 272], a latex protein defending *Morus* sp. trees from insect herbivory [211], a protein essential for tolerance to heat, salt and drought stresses in *Arabidopsis thaliana* [210], and a protein enhancing lignin accumulation in seedlings of *Arabidopsis* sp. [209]. The term "plant" is used to indicate Embryophyta, with two exceptions from the green algae *Klebsormidium nitens* in "loopful" chitinases. Eight homologous families include bacterial CHIT sequences. The main cluster contains the most characterized group of bacterial "loopless" chitinases (class IV bacterial chitinases, ID 5), two clusters contain sequences mainly from Proteobacteria species (IDs 6-7), and the other five are smaller homologous families (IDs 8 to 12) from different bacterial sources. It is interesting to note that not all "loopless" chitinases derive from Actinobacteria (> 90%), and a fraction come also from Myxococcales (> 3%), Firmicutes (> 1%),

Betaproteobacteria (> 1%) and Gammaproteobacteria (> 1%), enriched in species typically found in soils. The remaining five homologous families (IDs from 13 to 17) contain only a few tens of sequences from Fungi, Metazoa, and Oomycota, with the only fungal characterized seed from the homologous family with ID 14.

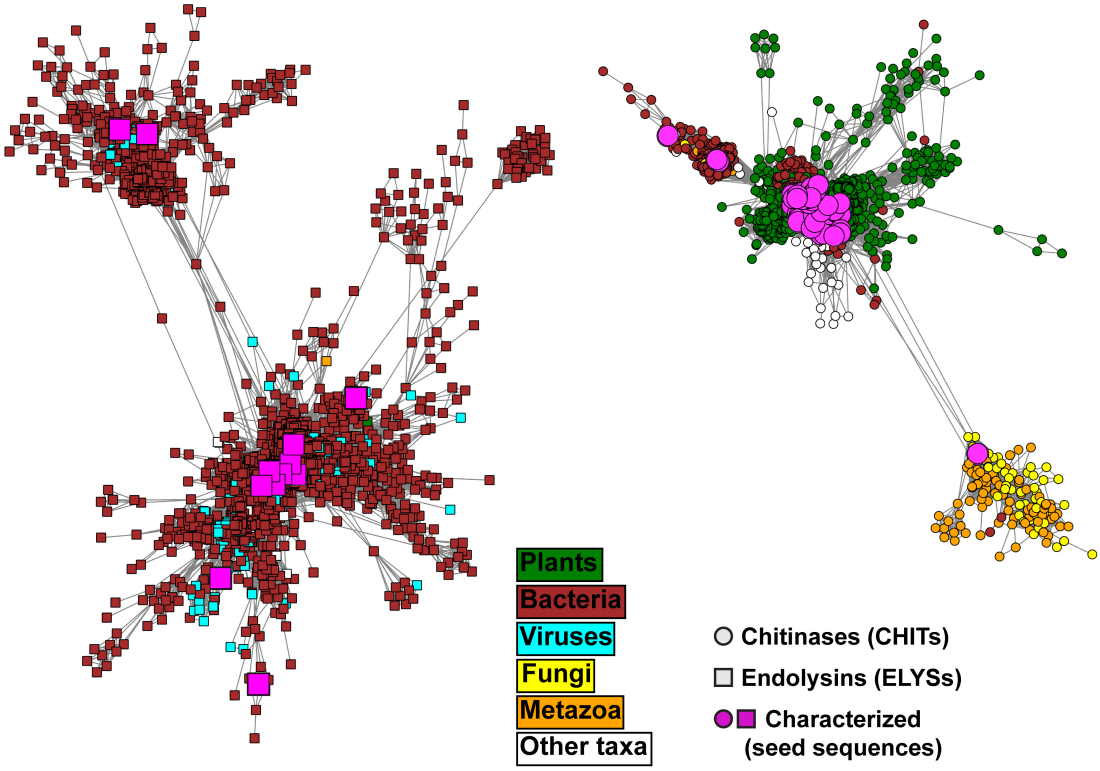


Figure 1. Protein sequence networks of representative domains of the two bigger clusters containing seed sequences (5067 centroid nodes, 2738 nodes on the left for ELYSs, endolysins, and 2329 nodes on the right for CHITs, chitinases) connected by edges with a sequence identity cut-off of 40%. The prefuse force-directed OpenCL layout was used for network visualization. The domains were extracted by scanning the sequences collected through BLAST searches (using the seed sequences reported in **Tab. S1** as queries) with Pfam's GH19 profile HMM (PF00182). Nodes are colored according to their annotated taxonomic source. The remaining smaller network clusters are visualized in **Fig. S3**.

The ELYS superfamily standard numbering scheme was created from the profile HMM obtained aligning the sequences of 12 biochemically characterized endolysins (**Tab. S1**). The endolysin from the bacteriophage SPN1S of *Salmonella typhimurium* [141], the only ELYS protein with a known structure (PDB accession 4ok7), was selected as reference. The standard numbering covers the catalytic domain from position 1 to 209 and was used to obtain networks of the catalytic domain centroids with nodes connected by edges if two centroids shared at least 60% sequence identity (**Fig. 2B**). Based on the resulting networks, ELYS sequences were assigned to 34 homologous families from bacteria or viruses (**Tab. S2**). Eight homologous families contain at least one characterized seed sequence and only two of

them contain thousands of sequence entries. One of these contains the reference seed endolysin from *Salmonella typhimurium* phage (ID 2), and the other one four seed endolysins from *Pseudomonas* phage/prophages (ID 1). Among other homologous families with more than 100 sequences, only one contains a *Mycobacterium* phage seed (ID 8), while the other five contain only uncharacterized putative endolysins. The remaining ELYS homologous families are small and contain just one sequence up to few tens at most.

The length distribution of both CHIT and ELYS domains is bimodal, more pronounced in CHITs (**Fig. S4**). The shorter length mode is around 200 for CHITs and 175 for ELYSs; the longer mode of CHITs is 245, whereas the longer mode for ELYSs is around 200, like the shorter mode of CHITs. Based on the number of amino acids in single domains, we hypothesize that within the sequence length distribution (**Fig. S2**), the shorter peak (around 200 residues) contains single domain proteins, whereas the longer one (around 580 residues) contains proteins organized in two or more domains. By looking into the sequences of annotated ELYSs and CHITs within this second peak (a window from 560 to 620 amino acids), 51% are represented by one CHIT domain in association with one CBM5/12 at the C-terminus and another putative, not yet characterized, domain at the N-terminus. A catalytic domain is associated with at least another undefined domain in another 40% of these sequences, while just a minority (6%) is formed by one CHIT domain and two CBM5/12 either at the N- or C-terminus. Only five sequences of that peak display one CBM18 at the N-terminus followed by two CHIT domains, or a CBM18 in between two CHIT domains. One sequence is made by two tandem blocks of CBM18 and CHIT domain, and just one sequence contains two CHIT domains.

2. Conservation analysis

From the multiple sequence alignments of the CHIT and ELYS superfamilies, the conservation of each standard position was derived through Rate4Site (see *Experimental Procedures*) and plotted on the reference structures (**Fig. S5**): 77 out of 242 positions in CHITs (**Tab. S3**) and 51 out of 209 positions in ELYSs (**Tab. S4**) had the highest conservation rate. Most (but not all) of the highly conserved positions were located in the active site cleft, while the least conserved residues were in the loops at the extremity of the catalytic cleft or at the surface of the two lobes (**Fig. S6**). The structural alignment of the highly conserved positions in CHITs (**Fig. 3A**-**C**) and ELYSs (**Fig. 3B**-**D**) highlights the presence of a conserved GH19 core spanning the catalytic centre and the internal part of each lobe. The conserved core comprises the catalytic residues (E69, E87, S120 and E49, E58, T130 for CHITs and ELYSs, respectively) and the substrate binding residues at subsites -2,-1, and +1 (H66, E67, E89, Y96, Q118, S120, N124, I198, N199, R215 and H48, E49, E58 Y106, Q128, T130, N134, I186, N187, R196 for CHITs and ELYSs, respectively). Other substrate binding positions at subsites -4, -3, +2, +3 and +4 were not conserved, neither in CHITs nor in ELYSs (**Tab. 1**). One structurally equivalent position predicted to bind the substrate at subsite +1 in the reference chitinase (standard positions E203 and N191 in CHITs and ELYSs, respectively) has the highest level of conservation, but it is not identified as part of the shared core since that position does not contain a gap in less than 90% of ELYS sequences. Thus, 27 positions are conserved among CHITs and ELYSs**,** whereas the pattern of residues in six positions in CHIT (97, 105, 151, 190, 192, and 222) and four positions in ELYS (33, 109, 118, and 173) was found to be present within the two superfamilies, but different between them (**Tab. 2**).

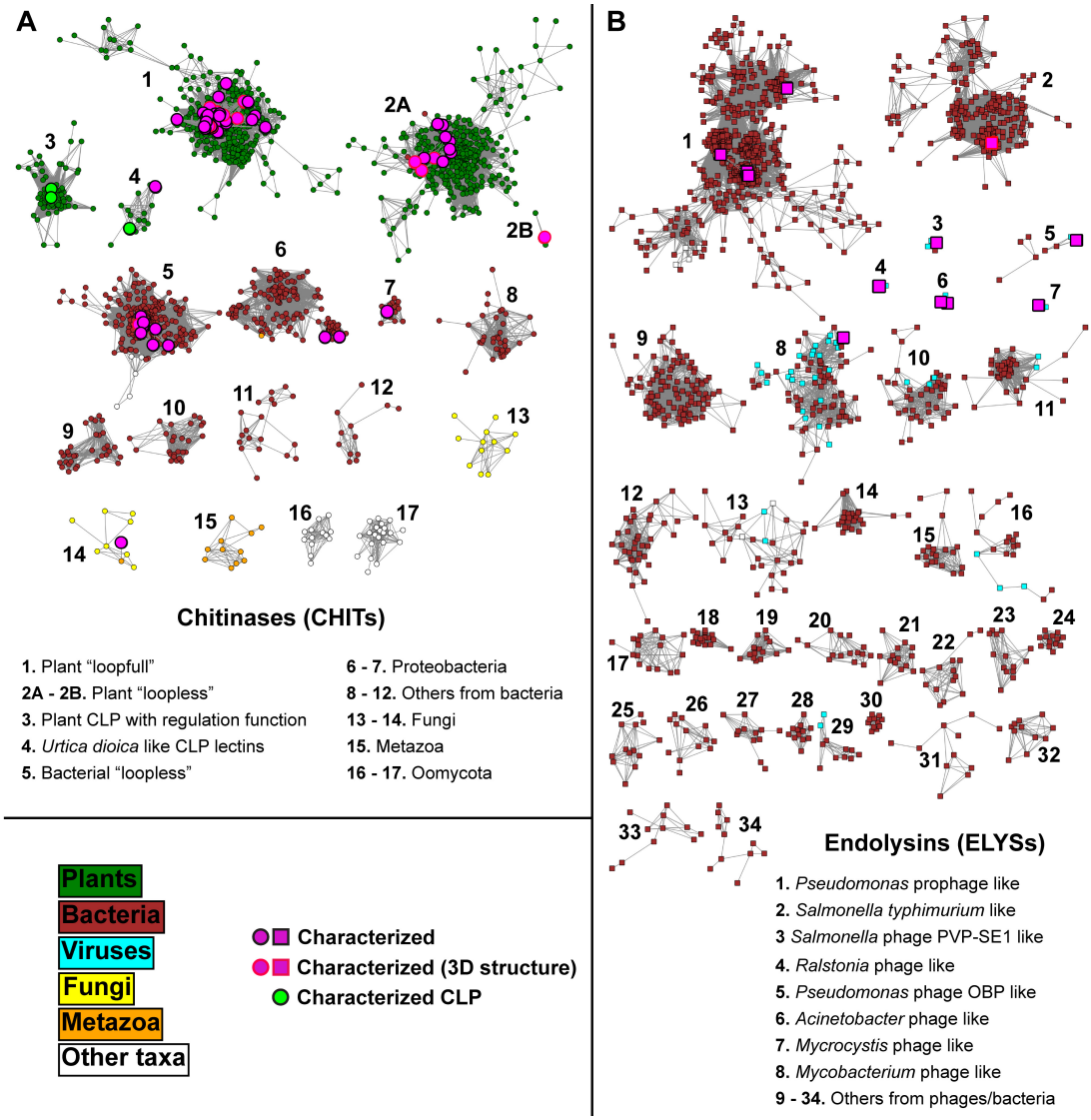


Figure 2. Protein sequence networks of representative domains of CHITs and ELYs (1860 centroid nodes for CHITs and 1521 centroid nodes for ELYSs, respectively) connected by edges with an identity cut-off of 60% sequence identity (used for homologous family assignment). The prefuse force-directed OpenCL layout was used for network visualization. The domains were extracted by using profile HMMs of CHITs and ELYSs (generated in this study) to scan the sequences in the GH19ED database. Nodes are colored according to their annotated taxonomic source. Seed sequences are highlighted, with a different border if a structure is available in the PDB. Nodes representing characterized "chitinase-like" proteins (CLPs) [132, 210, 211] are also highlighted.

Table 1. Conserved core shared in CHIT and ELYS superfamilies. Structurally aligned positions are listed in each row, numbered according to each superfamily-specific standard numbering scheme. Information is provided about the percentage of conserved residues, if higher than 5%. The reported function is indicated with respect to the reference chitinase from rye seed [139].

Standard position CHIT	Amino acid distribution CHIT			Standard position ELYS	Amino acid distribution ELYS			Function[a]
56	E 68%	S 23%	T 5.2%	38	R 60%	D 22%	W 4.5%	
58	A 80%	V 11%	I 4.2%	40	A 90%	S 2.1%	C 1.8%	
59	A 63%	T 32%		41	A 36%	M 34%	Y 9.3%	
60	F 61%	M 19%	A 16%	42	F 82%	W 4.5%	M 3.7%	
62	A 82%	G 12%	T 1.8%	44	A 91%	G 4.3%	S 3.4%	
63	H 52%	N 25%	Q 19%	45	Q 88%	T 9.3%		
66	H 54%	Q 34%	S 3.6%	48	H 93%	V 2.3%		Substrate binding (+1)
67	E 91%	K 4.7%		49	E 99%			Catalytic proton donor and substrate binding (-1)
68	T 91%	S 5.4%		50	S 86%	T 11%	C 1.4%	
89	E 94%			58	E 99%			Catalytic base and substrate binding (-1)
96	Y 90%	K 1.7%	M 1%	106	Y 95%	F 2.3%		Substrate binding (-1)
113	G 99%			123	G 99%			
114	R 98%	K 1.4%		124	R 95%	G 1.2%	A 1.2%	
115	G 99%			125	G 96%	T 2.5%		
118	Q 91%	P 4.8%	M 3.2%	128	Q 92%	M 5.5%	G 1%	Substrate binding (+1)
120	S 84%	T 9.6%	Y 4.3%	130	T 99%			Catalytic water coordination and substrate binding (-2)
124	N 99%			134	N 96%			Substrate binding (-2)
125	Y 99%			135	Y 97%	F 1.5%		
140	P 100%			150	P 95%	G 1.2%		
143	V 91%	I 4.2%	L 3.8%	153	L 74%	A 13%	V 7.8%	
154	A 85%	G 13%	S 1.3%	163	A 83%	S 4.2%	E 1.9%	
158	W 66%	F 30%	Y 2.3%	167	W 65%	F 14%	Y 10%	
195	I 63%	T 29%	M 4.3%	183	T 81%	R 9.6%	S 3.6%	
198	I 88%	L 8.2%	V 2.1%	186	I 85%	V 12%		Substrate binding without side chain (-2)
199	N 92%	Y 4.8%		187	N 97%			Substrate binding (-2)
200	G 94%	S 2.3%	A 1.8%	188	G 89%	L 3.3%	P 1.5%	
215	R 92%	I 4.3%		196	R 89%			Substrate binding (+1)

[a]Binding subsites (in parenthesis) are numbered according to the standard nomenclature; cleavage occurs between the sugar units bound at subsites -1 and +1 [140]

3. Loops in CHITs

The CHITs standard numbering scheme was applied to annotate the start and the end of each of the loops throughout the database (**Tab. 3**). The naming convention of the six loops is based on the comparison of "loopful" and "loopless" chitinases in the structural alignment of **Fig. 4A,** which resembles the definition reported in [207]. Loops 2, 3 and 5 are the longest (**Fig. S7**), and there is also a wide variation in length, except loops 4 and 5 that have a comparably narrow distribution.

Loop 4 is the most conserved in terms of sequence, while the loops 1, 2, 5, and the C-terminal are the most variable. The substrate binding sites located on loops are not conserved, except for standard position 96 on loop 2. Analysis of the length distribution showed that two different modes of length could be detected for loops 2 (from 10 to 16 residues and from 18 to 23 residues) and 3 (from 12 to 20 residues and from 22 to 31 residues). Longer loops 2 and 3 are found only in a Proteobacteria homologous family (ID 6).

Loop 3 was present in all homologous families but a small group of Proteobacteria CHITs (ID 7). The pattern of presence or absence of the other five loops varied between homologous families (**Tab. 4**). All six loops were present in the "loopful" plant CHITs (ID 1), plant chitinase-like protein (CLP) with regulatory functions (ID 3), and a small group of bacterial CHITs (ID 12), whereas other bacterial chitinases (IDs from 6 to 11) lacked at least loop 1, and the first 3 loops were absent in a Proteobacteria chitinase homologous family (ID 7). Loop 5 was absent in *Urtica dioica*-like CLP lectins (ID 4). The first three loops were present in most of the "loopless" plant CHITs (ID 2A-B), whereas loops 3 and 4 were present in the "loopless" bacterial CHITs (ID 5) and, with some variations, in CHITs from Fungi (IDs 13-14), Metazoa (ID 15), and Oomycota (IDs 16-17). The pattern of loops was summarized with a binary loop code presented in **Tab. 4.**

Table 2. Percentage distribution of amino acids at standard positions corresponding to the signatures specific for CHIT and ELYS superfamilies. Information is provided about the percentage of conserved residues at each superfamily-specific standard numbering scheme position, if higher than 5%.

CHIT Standard position	CHIT Amino acids distribution			ELYS Standard position	ELYS Amino acids distribution	
97	C 93%			33	I 80%	
105	C 91%			109	R 80%	E 5.4%
151	F 41%	L 32%	W 23%	118	G 97%	
190	G 95%			173	L 54%	Y 23%
192	G 95%					
222	F 39%	Y 34%	L 15%			

4. Accessory modules

The protein sequence networks of CHIT centroids together with the information about their association with known accessory binding modules are shown in **Fig. 5**. The catalytic domain of some bacterial chitinases (homologous families IDs 5, 6, 7, 9, and 10) are fused to CBM5/12, a few members of “loopless” bacterial chitinase (ID 5) are fused to CBM13, and some Cyanobacteria chitinase (in homologous family ID 11) are fused with LysM.

Most of the plant CHIT homologous families are associated with a CBM, except for plant CLP with regulatory functions (ID 3). Other eukaryotic and three distinct bacterial homologous families (IDs 8 and 11 to 17) do not contain any CBM. Remarkably, CHIT domains associated with a CBM were not clustered in a single region of the CHIT sequence space.

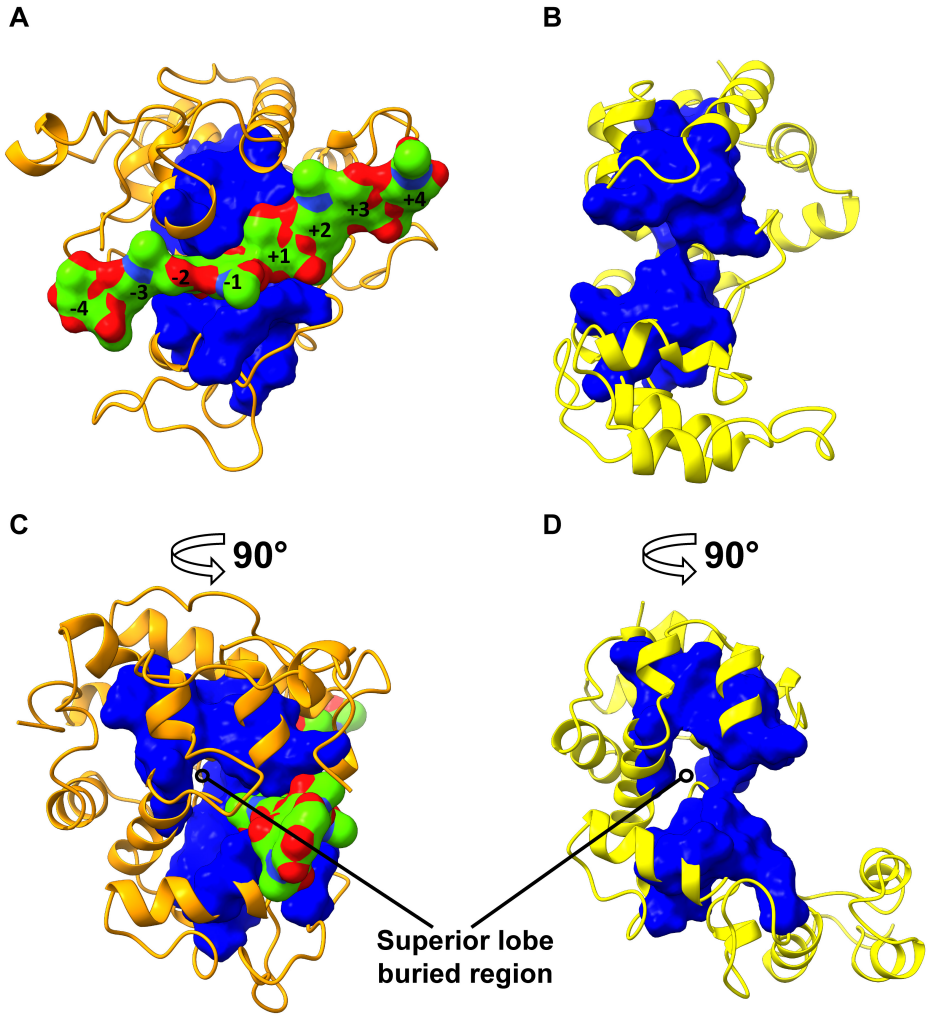


Figure 3. The most conserved and structurally aligned positions between CHITs and ELYSs (**Tab. 1**) are plotted in blue solvent accessible surface onto the reference models of CHIT (**A**) and ELYS (**B**) superfamilies (PDB accessions 4j0l and 4ok7, respectively), represented in cartoon style. In (**C**) and (**D**), the same models are rotated by 90° according to the vertical axis.

The protein sequence networks of ELYS domain centroids are shown in **Fig. 6** together with the information about their association with accessory binding modules. Peptidoglycan binding motifs but not CBMs were found in ELYSs (**Fig. 6**): LysM and PG_binding_1 are the only motifs present in more than 10 sequences throughout the ELYS superfamily. PG_binding_1 is the most spread domain, present in the sequences of two small homologous families (IDs 15 and 21) and in few sequences of other seven homologous families (IDs 1, 5, 12, 13, 14, 20, and 31). LysM can be found in most of the sequences of two small ELYS homologous families (ID 13 and 22) and few sequences of the biggest ELYS homologous family (ID 1)

In [141], the presence of a 3-helix peptidoglycan binding module (PBM) was reported as a new binding motif that covers standard positions 59 to 106 in the endolysin catalytic domain from bacteriophage SPN1S (**Fig. 4B**). In the Rate4Site conservation category assigned to these positions in

the ELYS superfamily, the average score is 1.8 (minimum is 1, maximum is 5), which means this region, like some of the chitinase loop motifs, can be considered as overall not conserved. In **Fig. 6** the sequences harbouring a PBM derive from the same homologous family of the reference endolysin (ID 2) from bacteriophage SPN1S of *Salmonella typhimurium* [141], and from another group (ID 19) which is small and contains endolysins from Enterobacteriales. PBM is also present in just few sequences with a *Pseudomonas* sp. background in *Pseudomonas* prophage like homologous family (ID 1), whereas all other ELYS sequences do not contain this motif.

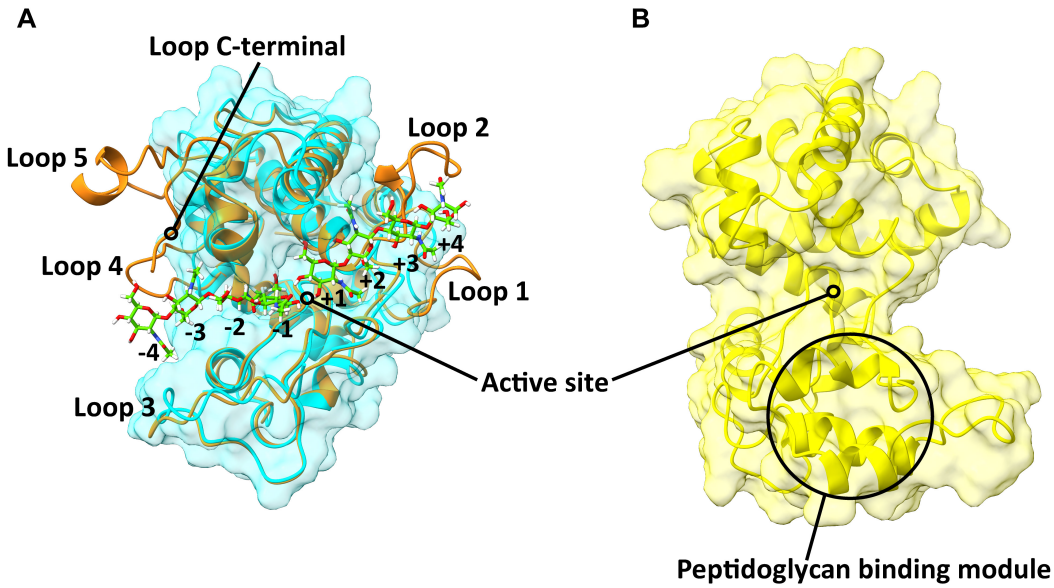


Figure 4. The structures of GH19 "loopful" chitinase from rye seed *Secale cereale* (orange, PDB accession 4jol [139]) and "loopless" chitinase from moss *Gemmabryum coronatum* (cyan, PDB accession 3wh1) superposed with the *mmaker* command implemented in ChimeraX 0.9 (**A**), showing the five additional loops of "loopful" plant chitinases. The two tetra-chitooligosaccharides spanning the catalytic cleft in complex with the crystal structure of rye seed are shown; numbers under sugar moieties are in accordance with the standard nomenclature for GH. Cleavage occurs between units bound in subsites -1 and +1 [140]. The structure of GH19 endolysin from bacteriophage SPN1S (PDB code 4ok7 [141]) of *Salmonella typhimurium* is shown for comparison (**B**).

5. Phylogenetic analysis of the GH19 family

A phylogenetic tree was built to study the evolutionary relationships between the GH19 enzymes, and to relate them to biochemical properties and structural patterns. In this analysis, we considered a comprehensive and representative sample of the GH19 sequence space by selecting 64 catalytic domains as clustered centroids from all CHIT and ELYS homologous families (**Fig. 7**). The result confirms that the homologous families of the two GH19 superfamilies have two distinct common ancestors. In the ELYSs branch, 34 out of 40 sequences are of bacterial origins, despite many biochemically characterized endolysins from phage or prophages.

The centroids of three different ELYS homologous families, whose IDs are 1 (*Pseudomonas* prophage like), 5 (*Pseudomonas* phage OBP like) and 16 (uncharacterized), seem not being closely related from an evolutionary point of view. The homologous families with IDs 2 (*Salmonella tiphymurium* like) and

19 (uncharacterized), which contain the PBM in their catalytic domain, share a common ancestor with homologous family ID 30 (uncharacterized).

It is not clear which type of organism most likely harboured the ancestral GH19 chitinase gene. The putative chitinase lineages of eukaryotic homologous families (Fungi, Metazoa and Oomycota) might have separated very early, before the radiation of the most successful groups of plant and bacterial chitinases, which share a more recent common ancestor. Bacterial "loopless" chitinases, the most characterized group of bacterial chitinases, separated before the evolution of plant "loopful" and "loopless" chitinases, which in turn seem to have the same common ancestor, but evolved along two distinct lineages. Moreover, the tree is compatible with the hypothesis that GH19 CHITs were transferred from plant to bacteria through two independent horizontal gene transfers, even if the posterior probabilities for the nodes corresponding to HGTs are not high (around 0.5 to 0.6). Interestingly, two centroid sequences from the same bacterial homologous family (ID 11) have different phylogenetic histories.

6. Properties derived from protein sequence networks

The 21851 sequence regions that correspond to the catalytic domains, or core regions, of CHIT, ELYS, and sequences not classified in a superfamily were aligned in an all-*vs*.-all approach as outlined for the protein sequence networks mentioned above. The distribution $N(n)$ of the degrees n, i.e. the distribution of the number of neighbouring sequences n, was approximated by a power-law function with a scaling exponent $\gamma = 1.1$ for $n \leq 50$ (**Fig. S8**). Thus, comparably few sequences of the core region were found to be densely connected to homologous neighbours, and an exemplary hub region in GH19 sequence space could be derived (**Tab. S5**). The histograms for the distributions of the number $N(s)$ of s sequences contained in a common cluster at thresholds of 60%, 70%, 80%, and 90% pairwise sequence identity could also be approximated by a power-law function (**Fig. S9**). The slopes τ_h of these distributions represent the ratios of small to large clusters and thereby indicate the connectedness of the GH19 sequence space, with an extrapolated exponent of $\tau = 1.1$ (**Fig. S10**)

Table 3. Loop conservation scores at CHIT standard positions. The "loopful" plant chitinase from rye seed (PDB accession 4j0l) is taken as reference. Conservation score ranges from 1 (least conserved) to 5 (most conserved). Substrate binding residues present in the reference chitinase are highlighted in bold.

Loop	Site	Conservation score	Average conservation score
1	20	1	2.1
	21	1	
	22	1	
	23	5	
	24	1	
	25	3	
	26	2	
	27	1	
2	**70**	**3**	1.2
	71	1	
	72	**1**	
	73	1	

	74	1	
	75	1	
	76	1	
	77	1	
	78	1	
	79	2	
	80	1	
	81	1	
	82	1	
	83	1	
3	94	1	2.71
	95	1	
	96	**5**	
	97	5	
	98	2	
	99	1	
	100	2	
	101	1	
	102	2	
	103	4	
	104	4	
	105	5	
	106	3	
	107	2	
4	160	4	3.9
	161	4	
	162	**4**	
	163	2	
	164	3	
	165	5	
	166	4	
	167	5	
	168	4	
5	174	3	1.1
	175	1	
	176	1	
	177	1	
	178	1	
	179	1	
	180	1	
	181	1	
	182	1	
	183	1	
	184	1	
	185	1	
	186	1	
	187	1	
	188	1	
C terminal	236	5	1.5
	237	1	
	238	1	
	239	1	
	240	1	
	241	1	
	242	**1**	
	243	1	

Discussion

1. Extended classification of GH19 sequence space

Our study showed that, based on the distance in the sequence space, biochemically characterized GH19 chitinases and endolysins (retrieved from literature and presented in **Tab. S1**) can be separated into two superfamilies, the CHITs and the ELYSs. CHITs consist of four large (>1000 sequences) homologous families, 1 of intermediate (100-1000 sequences) size, and 12 of small size (<100 sequences). ELYSs consist of two large (>2500 sequences) homologous families, 4 of intermediate (100-1000 sequences) size, and 26 of small size (<100 sequences). Therefore, if looking to the sequence space distribution, it is possible to note a difference between ELYSs and CHITs: in ELYS superfamily most of the sequences are distributed on two very big homologous families and the rest is sparse in small homologous families which are more than double than the ones present in CHIT superfamily. This result can be responsible for the limited significance of the inferred scale-free properties of GH19. However, we argue that this discrepancy is a consequence of dealing with a relatively high (>50%) number of putative endolysin sequences from phages or phage-related regions in bacteria genomes: indeed, phages are the most abundant and diverse self-replicating entities in the planet [273] and only a part of this diversity is more densely studied because of its importance in human health; this can explain the difference in ELYS sequence space, if compared to CHITs.

Seven CHIT homologous families and eight ELYS homologous families contain at least one characterized enzyme. Anyway, most of the characterized GH19 are CHITs (63), mainly from plant sources (49), whereas characterized ELYSs (12) are phage or prophage endolysins, although most of ELYSs come from bacteria, suggesting that phage sequences may have been internalized in the genome of the bacterial host, and assigned to the host species in NCBI databases.

The distribution of loop-based classification across the entire sequence space of previously defined classes was not established in previous work [122]; therefore we first established homologous families based on clusters obtained by pairwise comparison of GH19 catalytic domains, and then used them in order to make a first comprehensive comparison and extension of these systems of classification: plant class I and II chitinases ("loopful") were assigned to a single plant homologous family (ID 1), and class IV chitinases ("loopless") was separated into two homologous families (plant: ID 2A-B, bacteria: ID 5). Two new plant homologous families (ID 3, 4) contained GH19 proteins, which were characterized as mediators of plant growth (ID 3) or catalytically inactive lectins (ID 4). Therefore, the other sequences in these two homologous families are predicted as putative CLPs. In addition, we introduced 12 additional homologous families from bacteria (ID 6-12), Fungi (ID 13,14), Oomycota (ID 16,17), and Metazoa (ID 15). The sequences in the small bacterial homologous family 7 lack the N-terminal catalytic part, which contains the first 3 loops. However, a biochemically characterized member of this family has been described as an active chitinase hydrolyzing hexachitooligosaccharides at the second bond from the non-reducing end [128], with a high free energy of binding at subsites +3 and +4 [274], whereas most of other plant GH19 chitinases have higher affinity for binding at subsites from -3 to +3 [129]. Thus, we predict that the members of homologous family 7 preferably hydrolyze substrates at the non-reducing end. Interestingly, the same selectivity was found for members of chitinase “loopful” homologous family, where a N-terminal region comprising loop 2 was suggested to mediate this function [201, 218, 220]: therefore, we think that also the members of homologous family 7 contain an N-terminal region with a similar function, but non-homologous to other plant “loopful” CHITs.

Three plant CHIT homologous families (IDs 1-2-4) included members with a fused CBM18 domain, despite it is not present in all the sequences and does not fit any known sequence classification pattern by looking to the sequence space. Only the smallest plant homologous family (ID 3) did not possess any CBM; the same scenario applies for CBMs associated to chitinases in bacterial homologous families. Therefore, we suggest that the presence or absence of CBMs should not be used as a main criterion for the classification of GH19 diversity in chitinases; however, CBMs are usually associated to an improved antifungal activity [262], thus it is suggested to look for the presence of accessory binding modules when selecting GH19 sequences required to have this property.

Table 4. Percentage distributions of loop annotations among CHIT homologous families. Names are defined according to **Fig. 2A** (occurrences not displayed if below 5%). h-fam = homologous family name; ID = homologous family identifier. Loop code: '0' = absent; '1' = present; '-' = undefined.

CHIT h-fam (ID)	Loop 1	Loop 2	Loop 3	Loop 4	Loop 5	Loop C-terminal	loop code
Plant "loopful" (1)	88.2%	93.1%	88.7%	97.8%	96.7%	91.8%	1 1 1 1 1 1
Plant "loopless" (2)	95.4%	89.4%	99.8%	5.6%			1 1 1 0 0 0
Plant CLP with regulation function (3)	90.5%	95.5%	97.0%	100%	99.8%	94.5%	1 1 1 1 1 1
Urtica dioica like CLP lectins (4)	96.8%	96.8%	100%	100%%		71.0%	1 1 1 1 0 1
Bacteria"loopless" (5)			99.9%	99.8%			0 0 1 1 0 0
Proteobacteria (6)		51.7% [a]47.4%	[a]99.1%	99.7%	99.6%	98.0%	0 1 1 1 1 1
Proteobacteria (7)				100%	100%	93.3%	0 0 0 1 1 1
Bacteria (8)		97.0%	98.5%	100%	98.5%	76.5%	0 1 1 1 1 1
Bacteria (9)		100%	100%	100%	100%	78.6%	0 1 1 1 1 1
Bacteria (10)		100%	100%	100%	100%	15.1%	0 1 1 1 1 0
Bacteria (11)	17.9%	53.6%	100%	78.6%	60.7%	42.9%	0 - 1 1 - -
Bacteria (12)	93.0%	100%	100%	100%	100%	100%	1 1 1 1 1 1
Fungi (13)			100%	73.3%		66.7%	0 0 1 1 0 2
Fungi (14)		30%	60%	88.9%			0 0 - 1 0 0
Metazoa (15)			100%	22.2%			0 0 1 0 0 0
Oomycota (16)			95.6%	98.5%			0 0 1 1 0 0
Oomycota (17)	19.6%		100%	97.8%			0 0 1 1 0 0

[a]This fraction of sequences have a longer loop, based on length distribution reported in **Fig. S7.**

2. Signatures of substrate specificity

Structure-based comparison done in other studies suggested that the catalytic residues and the central substrate binding subsites -2, -1 , +1 have been described as conserved in GH19 chitinases [111, 125, 139, 201, 213, 215, 216]. We found that 27 CHITs residues, among which the ones corresponding to this region, are conserved also in ELYSs, while the residues exposed on other subsites are variable both in CHITs and ELYs. This result can explain why in GH19 some CHITs and a single ELYS can bind and hydrolyze murein and chitin, respectively, despite the structural differences between the two substrates. In [141], a GH19 endolysin structure was superposed to structures of enzymes from other

families in the lysozyme superfamily, concluding that it is functionally an N-acetyl-β-D muramidase because the position of catalytic acid and base residues is compatible to C-type lysozyme (GH22). However, GH22 enzymes have a retaining catalytic mechanism with the involvement of acetamido substituent at C2 of the substrate in the catalysis [106].

Instead, we suggest the mechanism of hydrolysis of GH19 ELYSs to be similar to the GH19 CHITs, based on our conservation studies and the fact that these are two closely related superfamilies. Despite this, ELYSs differ from CHITs by shorter loops and a larger substrate binding cleft at subsites from -4 to +3 (also if only loops 3 and 4, present in the common ancestor of CHITs, are considered) to accommodate the more bulky murein substrate (**Fig. S11**) and possibly by a different opening or closing dynamic of the cleft during catalysis.

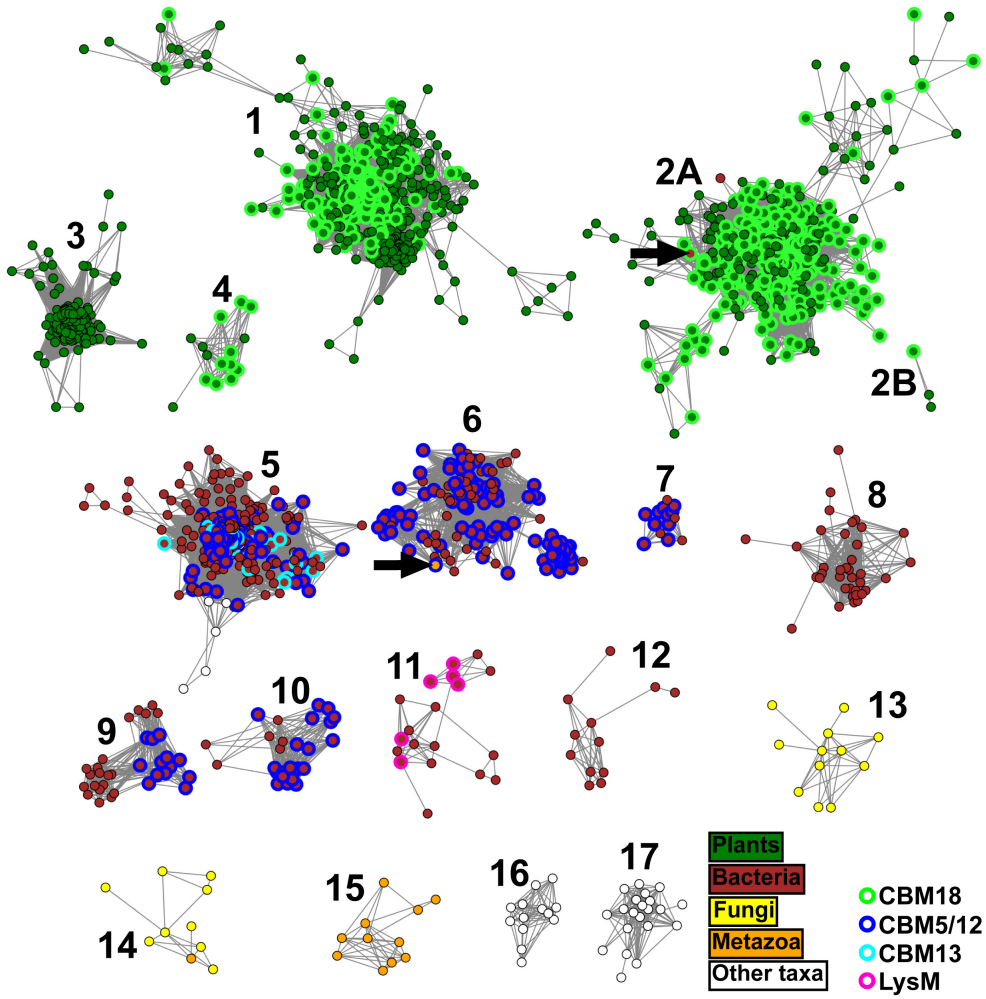


Figure 5. The presence of accessory binding modules is plotted with different colors onto CHIT homologous families sequence networks. The two black arrows indicate the centroids from the bacteria and Metazoa possessing a CBM18 (typical of plants) and a CBM5/12 (typical of Bacteria), respectively (see main text for details). The homologous family identifiers are the same as in **Fig. 2A**..

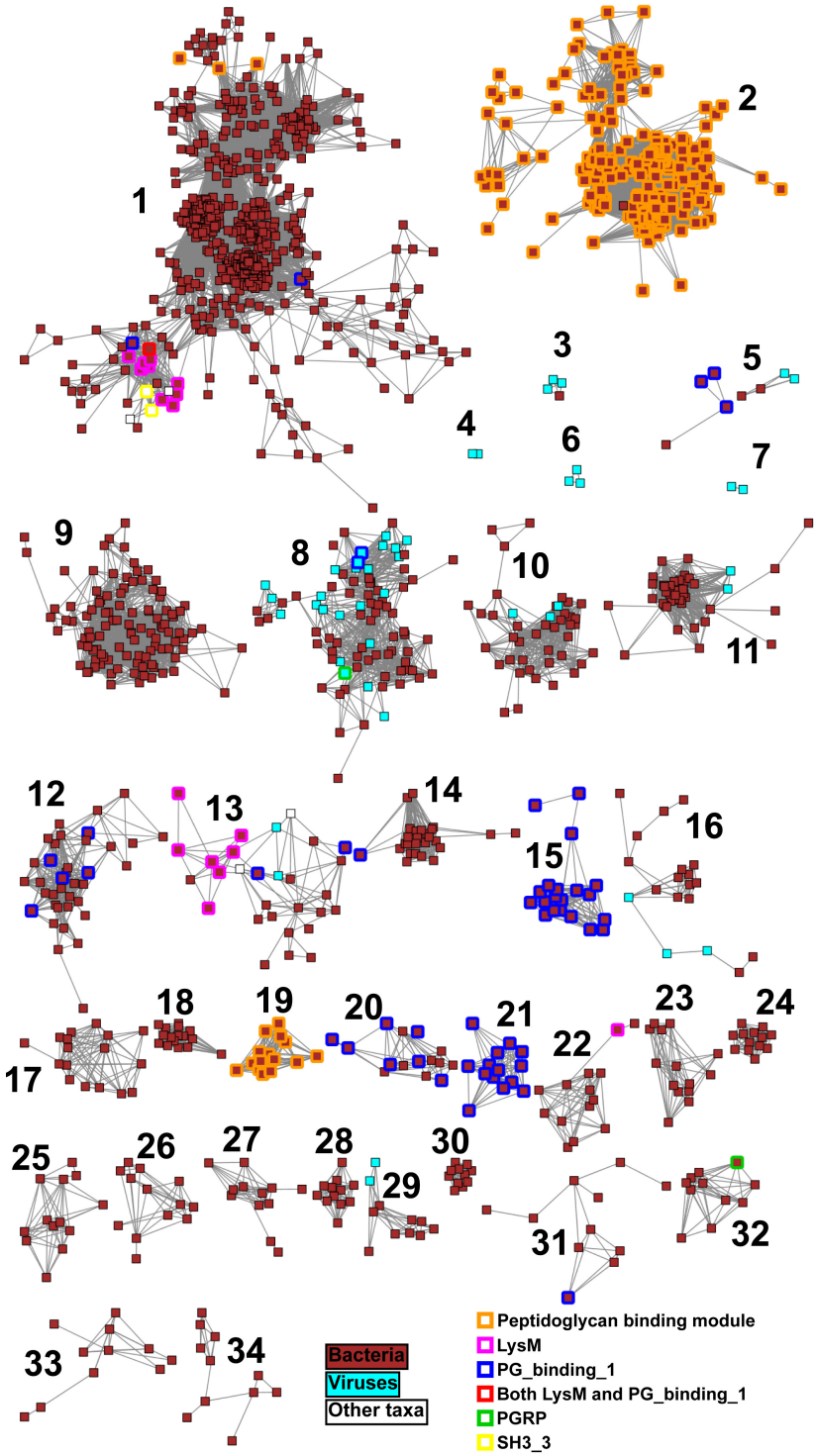


Figure 6. The presence of accessory binding modules is plotted with different colors onto ELYS homologous families sequence networks. The homologous family identifiers are the same as in **Fig. 2B**.

The most common residues present in positions conserved specifically in CHIT and ELYS sequences (**Tab. 2**) provided a basis for the identification of a specifically conserved signature which might mediate specificity toward chitin and murein, respectively. The signature coding for chitin hydrolysis is made by a distribution of residues over six positions, four of which are found in sixty-two of sixty-three characterized CHITs. The only outlier is a GH19 chitinase characterized from a fungus species [223], located far from the other characterized CHITs, in the lower-right portion of the CHIT network in **Fig. 1**. At least three positions of the four residues signature coding for murein hydrolysis was found in ten of twelve characterized ELYSs. The two outliers are two endolysins from *Acinetobacter* phages, located far from the other characterized ELYSs, in the upper-left portion of the ELYS network in **Fig. 1**: they possess a unique C-terminal amphipathic helix, predicted to facilitate the permeabilization of Gram-negative outer membrane [112, 134]. In the CHIT reference structure, the residues corresponding to the signature (**Fig. 8A-C**) are two cysteines forming a disulfide bridge potentially controlling the rigidity of loop 3, a phenylalanine located behind the active site in the hinge between the lobes, two nearby glycine residues, and a tyrosine inside the superior lobe over the catalytic cleft, with a possible role in flexibility control during the reaction. In the ELYS reference structure, the signature (**Fig. 8B-D**) is made by a methionine and an isoleucine residue far apart along the sequence, but interacting in the posterior part of the superior lobe of the structure; the other two residues are an arginine and a glycine between the hinge and the PBM. Because the residues of the two signatures are not in contact with the substrate, we argue that they participate in substrate specificity probably by mediating conformational changes upon substrate binding, which was observed in structural studies on other bilobal glycosidases [111, 139, 222, 275]. Unfortunately, no variants have been studied at these positions, yet.

3. GH19 evolution: loops, catalytic domain and accessory binding modules

GH19 evolution towards lysozyme (for ELYSs) and chitinolytic specializations (for CHITs) started very early. ELYSs remained confined in the genomes of phages and their bacterial hosts, specializing as lysozyme, without a noteworthy increase or decrease in sequence length, while chitinases spread also on eukaryotic taxa, increasing in length by the addition of loop insertions.

If the six different loops were distributed randomly among sequences, we should expect $2^6 = 64$ combinations of absent/present loop codes (in which 0 stands for “absence” and “1” for “presence”). However, only 8 combinations were found in the majority of sequences per each homologous family: 111111 (IDs 1,3,12), 111000 (ID 2), 111101 (ID 4), 001100 (IDs 5,16,17), 011111 (IDs 6,8,9), 000111 (ID 7), 011110 (ID 10), 001000 (ID 15), indicating that loops were added and lost according to a certain order, that can be inferred by looking into the details of CHIT evolution (**Fig. 7**). The CHIT ancestor included the most conserved loops 3 and 4 (loop pattern: 001100). This "loopless" CHIT remained of the same length over eukaryotic lineages, while at a certain point it split between a lineage of exclusively bacterial chitinases, resulting in the bacterial CHITs, and a lineage of plant chitinases in which loops 1 and 2 were added. Subsequently, plant CHITs diversified into “loopless” that lost loop 4 and into “loopful”, by addition of loops 5 and C-terminal. Bacterial CHITs further diversified in sequence and function, dividing between a lineage of Cyanobacteria (part of homologous family 11) and “loopless” chitinases of organic matter processing bacteria inhabiting the soils may have played an important role to in their ecological niche. Plant chitinases further diversified probably because of their importance in the plant immune system. Two derived groups of “loopful” plant CHITs, from

homologous families that likely lost enzymatic activity (ID 3 and 4), became coagulant factors in latex or plant growth mediators, in response to changing environmental conditions. Two late HGTs events were inferred to have moved GH19 from plants secondarily to many different taxonomic groups of bacteria, not only typical organic matter degrading strains: therefore, as just three GH19 enzymes were described in these groups of organisms, they would require more attention in future experimental studies. This is particular true for members of homologous family 7, which have a modified N-terminal region, as described above, and the homologous family 6, as loops 2 and 3 became longer, and evidence of exo-activity was experimentally collected in the two characterized GH19 chitinases of this group, from *Vibrio proteolyticus* and *Pseudoalteromonas tunicata* [129, 130]. GH19 CHITs are typically endo-acting enzymes. Thus, the appearance of longer loops during the evolution of homologous family 6 may explain the processive exo-activity, as supported by observations on other GHs, where processivity was considered as the result of longer loops responsible for the conversion of the active site shape from a cleft to a tunnel [95]. It should be noted that a few sequences from plant and bacterial "loopless" homologous families (IDs 2 and 5) possess a longer loop 2, and that two sequences from a bacterial homologous family (ID 10) possess a longer loop 3, thus demonstrating the broad diversity of possible loop patterns that were explored during evolution, although the substrate specificity of the respective gene products is yet unknown.

ELYSs sequences are shorter than CHITs and did not show distinct loop patterns [141]. Instead of loop 3 in CHITs, some ELYSs have a peptidoglycan binding module (3-helix bundle in case of the reference endolysin from *Salmonella* phage SPN1S). However, this module is not conserved among ELYSs, but only exists in two closely related homologous families (ID 2 and 19). Therefore, it is likely that it has recently evolved from an insertion under the selective pressure of co-evolutionary phage-bacteria interaction process, which is a key factor in increasing the rate of molecular evolution [196].

Many CHITs are linked to accessory domains. In plants, the prevalence of CBM18 can be explained by its presence in the early ancestor, whereas in Bacteria CBM5/12 was combined with CHITs after horizontal gene transfer from plants. A few members of the "loopless" bacterial CHIT homologous family are linked to CBM13, which is associated with Actinobacteria xylanases and is frequently present in multi-domain enzymes [276]. Therefore, we hypothesize that this domain was added recently to CHITs, as well as LysM, which is a ubiquitous non-catalytic motif repeat that was shown to bind both peptidoglycan and chitin in plants [224], and was mainly found in a subgroup of "loopless" bacterial CHITs from Cyanobacteria (ID 11). Two exceptions in the taxonomic distribution of CBMs were observed: a bacterial sequence in the plant "loopless" chitinases (ID 2A) harboring a CBM18, and a metazoan chitinase in a bacterial family (ID 6) harboring a CBM5/12 (indicated in **Fig. 5** by black arrows). We hypothesize that for these sequences both CBM and catalytic domain were transferred to these organisms from plants and bacteria, respectively. In ELYSs, only two known accessory binding modules are present (PG_binding_1, which is a peptidoglycan binding domain consisting of a three-helical bundle with a left-handed twist, and LysM) in a few hundred sequences (including more than 10 centroid sequences), which are spread among different homologous families with low sequence similarity. This finding supports the hypothesis of multiple independent recombination events with the ELYS catalytic domain.

Less than 50% sequences in the second peak of sequence length distribution (**Fig. S2**) have all their domains annotated. Therefore, further accessory domains, catalytic or not, conferring yet unknown and potentially interesting functions could be present in CHITs and ELYSs.

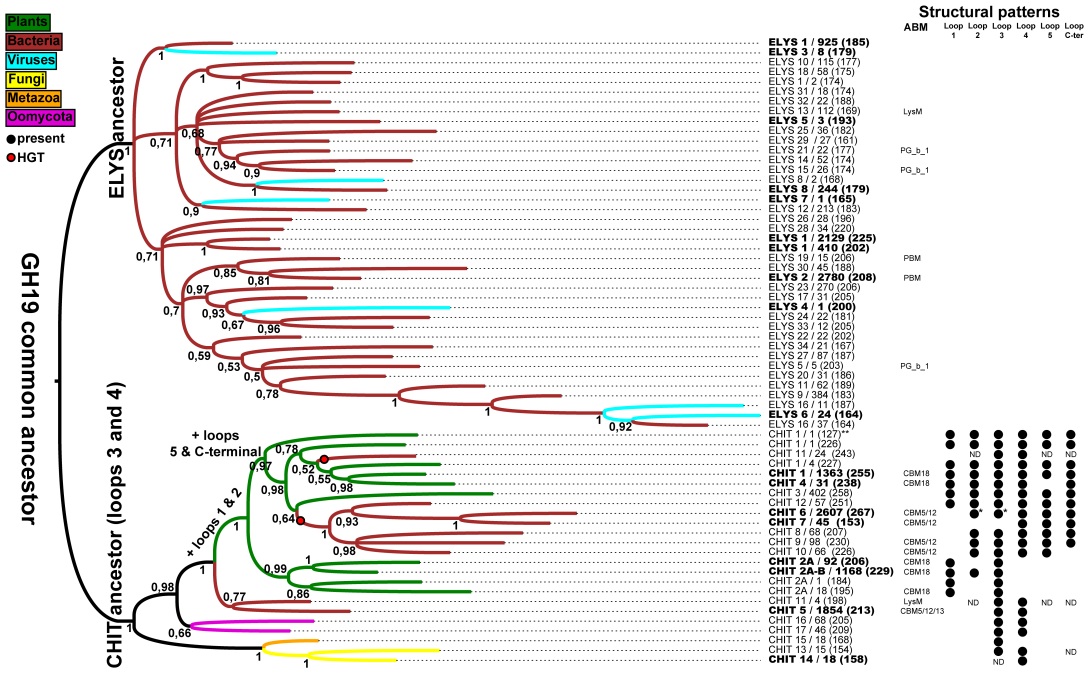


Figure 7. Phylogeny of centroids representative of GH19 sequence space, plotted on structural patterns analyzed in this study. Sequences are indicated with the respective superfamily name (ELYS or CHIT) followed by the homologous family identifier / n° of represented sequences, followed by sequence length in parentheses. Sequences representing characterized seed sequences are depicted in bold. *A fraction of sequences possess longer loops (see **Tab. 4**). **This sequence is a fragment. HGT = horizontal gene transfer; ABM = accessory binding module; PG_b_1 = PG_binding_1; CBM = Carbohydrate binding module; ND = Not defined because too variable. Posterior probabilities are indicated at nodes, which were collapsed and considered not informative if less than 0.5.

ELYS 1 = *Pseudomonas* prophage like; ELYS 2 = *Salmonella typhimurium* like; ELYS 3 = *Salmonella* phage PVP-SE1 like; ELYS 4 = *Ralstonia* phage like; ELYS 5 = *Pseudomonas* phage OBP like; ELYS 6 = *Acinetobacter* phage like; ELYS 7 = *Mycrocystis* phage like; ELYS 8 = *Mycobacterium* phage like; ELYS 9 to 34 = other putative endolysins from phages and prophages; CHIT 1 = plant "loopful"; 2A-B = plant "loopless"; CHIT 3 = plant CLP with regulation function; CHIT 4 = *Urtica dioica* like CLP lectins; CHIT 5 = bacterial "loopless"; CHIT 6-7 = Proteobacteria; CHIT 8 to 12 = other putative chitinase from Bacteria; CHIT 13-14 = Fungi; CHIT 15 = Metazoa; CHIT 16-17 = Oomycota.

4. Conclusion and outlook

In conclusion, we retrieved and analyzed more than 20000 sequences, containing at least a GH19 domain, and organized them in superfamilies that correlate with the chitinolytic and lysozyme activity specializations detected in experimental works. Moreover, GH19 superfamilies were classified in homologous families, extending the literature classification and providing a public database (GH19ED) for storing, annotating and organizing sequences, structures, domains and experimental references. Standard numbering schemes were developed that will help to pass functional annotations to

sequences still uncharacterized and map the identified signatures coding for chitin and murein hydrolysis. A loop code for fast description of GH19 chitinase loops was developed that may be correlated with specific catalytic properties, and we proposed a new evolutionary hypothesis of GH19 homologous families, which is different from past reconstructions in which plant sequences were overrepresented in the sample [114, 126], and explains the observed plasticity of structural elements, accessory domains and biological roles in different organisms. The actual lack of attention on non-plant GH19s was also mentioned in one of the last chitinase review [104]. In this regard, our work will contribute to the search for new interesting GH19 candidate enzymes in the homologous families from which no sequence has been described yet, especially most of the putative endolysin clusters and chitinases from bacterial and eukaryotic taxa. Moreover, we have built the necessary framework to disentangle the molecular basis of GH19 promiscuity, by guiding rational-based design of mutations. The capacity to predict and exploit new GH19 functions and properties will benefit also from future efforts into biochemical and structural characterization of still unknown and unannotated accessory domains, especially for GH19 endolysins.

Experimental Procedures

1. GH19ED database setup

In order to select the sequences for the database creation and subsequent analysis presented in this study, BLAST [37] searches were performed using as query the catalytic domain of a list of seed sequences against the NCBI non-redundant protein database (https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/)[277] and the Protein Data Bank (PDB, https://www.rcsb.org/)[278] with a maximal E-value of 10^{-10}. The seed sequences were obtained by selecting GH19 sequences stored in CAZy (http://www.cazy.org/GH19.html) [79] and Uniprot (https://www.uniprot.org/)[279], both accessed on 01/05/2019, and for which a biochemical characterization was found in literature.

Retrieved sequences, their label and organism source were inserted into GH19ED database (https://www.gh19ed.biocatnet.de) within the BioCatNet database system [227].

A global sequence identity threshold of 99% was applied to assign individual sequence entries to summarized *protein* ontology in the database.

2. Protein sequence networks

The sequence space was explored by obtaining protein sequence networks created from GH19 domain-level sequences, after being sorted with decreasing sequence length and clustered by 90% global identity using the heuristic clustering algorithm of USEARCH (UCLUST) [280]. The resulting reduced list of representative domain-level sequences, named centroids, is used in order to reduce the computational efforts for pairwise sequence alignments, performed with the Needleman-Wunsch algorithm implemented in the EMBOSS software suite [281], by the use of the BLOSUM62 scoring matrix [282] and maintaining defaults for gap opening and for gap extension penalties (10 and 0.5, respectively). The GNU parallel [283] package was employed to reduce the computational time for all the pairwise alignments by multithreading. Pairwise sequence identities between aligned centroids were calculated. Sequence networks were generated considering the centroids as the nodes and the

values of pairwise identity as the edge weights between each pair of nodes, with a certain minimum cut-off. Networks were then visualized with Cytoscape 3.7.1 [208] using the prefuse force-directed OpenCL layout algorithm with respect to the edge weights (i.e. sequence pairs with higher sequence identity tend to be placed in closer proximity).

The number of neighbouring nodes n for a given node, determined at a predefined threshold of sequence identity, is known as the degree of a node. The number of nodes $N(n)$ having a degree of n was fitted by a power law $N(n) \sim n^{-\gamma}$, and the scaling exponent γ was derived from a log-log plot [284].

Thresholds of global sequence identity were applied to form clusters of homologous sequences. The number of nodes $N(s)$ of cluster size s (with s being the number of sequences in a given cluster) was fitted by a power law $N(s) \sim s^{-\tau}$, and the Fisher exponent τ was determined as slope in a log-log plot, too [285]. Logarithmic histograms were formed for subsequent intervals between $s \geq 2$ and $s \leq 10$, between $s \geq 11$ and $s \leq 100$, $s \geq 101$ and $s \leq 1000$, and between $s \geq 1001$ and $s \leq 10{,}000$, respectively. The slopes τ_h of these histograms were determined at different thresholds of sequence identity. The slopes τ of the actual distribution were approximated by fits of τ_h against a power-law model distribution as described previously [285].

Distributions of degrees and cluster sizes were analysed by linear fitting via the fitlm function from the Statistics and Machine Learning Toolbox (version 11.5) in MATLAB (version R2019a, The MathWorks, Natick, MA, USA).

3. Superfamily assignment

The GH19 profile HMM from Pfam (PF00182) [286] was used for scanning the sequences contained in the database using the HMMER software suite (Version 3.1b2) [287]. A list of domain-based sequence hits was retrieved, and the sequences without at least 1 hit were removed from the database. The scanning parameters used to define domain hits in HMMER were a maximal E-value of 10^{-5}, a minimum hit length of 120 residues and a bias ratio (HMMER bias/HMMER profile-sequence alignment score) < 0.1 (the latter criterion was chosen to reduce false positives due to an uneven amino acid composition). The list of retrieved domains was used to generate protein sequence networks based on pairwise identities with the approach described in the above paragraph. By plotting the characterized seed sequences into networks at 40% identity minimum cut-off, centroid sequences belonging to different clusters (and the sequences they represent after previous USEARCH clustering) were assigned to a certain superfamily ontology entry within the GH19ED. Sequences that were found to be not connected to networks in which there were not seed sequences with available experimental data were included in the database, but not analyzed further.

4. Standard numbering schemes

The GH19 superfamily networks comprising at least one seed with a known crystal structure from the PDB, were refined by the creation of a superfamily specific standard numbering scheme according to the definition presented in [288]. For this purpose, a sequence with a known PDB structure was chosen as reference, along with a profile HMM derived from a multiple sequence alignment between that reference and other sequences. A starting alignment was built if other sequences with a known PDB structure were available, by performing a GH19 domain structure-based alignment generated through the *mmaker* command implemented in ChimeraX 0.9 (RBVI, University of California, San

Francisco, CA, USA) [289]. Other seed sequences in the same superfamily were added to this fixed structural alignment by the use of "--add" flag option available in MAFFT 7.407 [240] (described under https://mafft.cbrc.jp/alignment/server/add.html). If no other structures were available rather than the reference, a sequence-based alignment with other seeds was created with MAFFT "L-INS-i" strategy [290], improved by adding information of up to 600 close homologs obtained from a search in Uniprot non-redundant Uniref50 database (ftp://ftp.uniprot.org/pub/databases/uniprot/uniref/uniref50) using a restrictive E-value threshold of 10^{-20} (a procedure described more in detail at https://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html#homologs). The obtained alignments were manually cut with respect to the length of the GH19 domain of the reference structure and used to generate the new superfamily-specific profile HMMs for the assignment of standard position numbers.

5. Homologous family assignment

The standard numbering schemes obtained as described in the previous paragraph were used to annotate the GH19 domains in the sequences of the respective superfamily. The domain-level sequences were retrieved and aligned to calculate pairwise identities and build networks as outlined in paragraph 2. A 60% identity cut-off was used to split each superfamily into clusters (sub-networks) in order to better represent and organize the diversity in the sequence space. Each cluster containing at least one seed sequence or formed by more than 10 centroid sequences were assigned to a homologous family ontology entry. The available classification of GH19 (see "Introduction" section), the properties of seed sequences and the taxonomic source of sequence majority was considered for homologous family nomenclature.

6. Conservation analysis

A conservation analysis was made on superfamilies for which a profile HMM and a standard numbering scheme with respect to a reference sequence were obtained. For each superfamily, the domain-level sequences were clustered in descending length order with USEARCH as mentioned above, but employing a 65% cut-off identity in order to pick up less than 300 representative centroids (comprised the reference sequence) for generating a multiple sequence alignment with the E-ins-I algorithm of MAFFT [290]. Rate4Site (Version 2.01, [247]) computes the relative evolutionary rate at each site with respect to a multiple sequence alignment. Rate4Site rate scores were computed, normalized, and split in five different categories from the "fastest" (assigned to 1) to the "slowest" (assigned to 5). The multiple sequence alignment obtained with MAFFT was used as input with an LG substitution rate matrix [243] and an empirical Bayesian approach. If half or less sequences in the alignment contained gaps for a specific site, the "fastest" evolving score was assigned to the corresponding alignment column. Results were plotted on standard positions of each superfamily reference sequence and structure. The most conserved positions (Rate4Site category 5) without a gap in at least 90% of all sequences were identified in each superfamily. The structurally aligned positions of superfamily reference sequences collected from the previous step were selected as the shared GH19 "core", only if at those positions the amino acid frequency distribution in different superfamilies overlaps for more than 5%. On contrary, standard positions structurally aligned or not and overlapping

for less than 5% in the amino acid frequency distribution of different superfamilies, were considered specifically conserved within each superfamily. All the images of the mapped residues onto models were prepared with ChimeraX 0.9. Literature on reference sequences was employed in order to annotate known functions at standard positions.

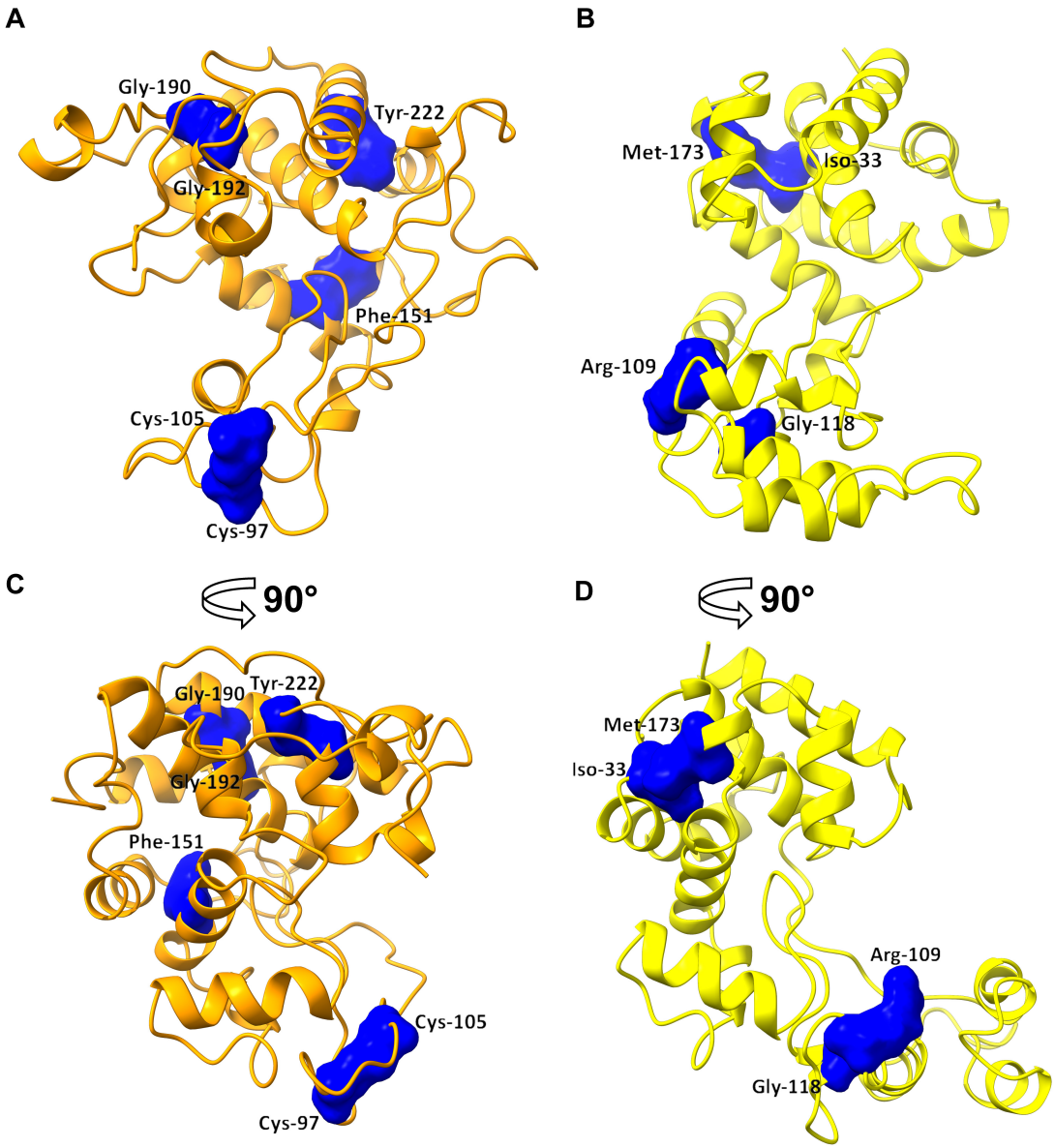


Figure 8. The residues corresponding to the CHIT (**A**) and ELYS (**B**) superfamily sequence signatures identified in this work are labelled and shown as blue solvent accessible surfaces onto the reference models (PDB accessions 4j0l and 4ok7 for CHIT and ELYS superfamily, respectively), displayed in cartoon style. In (**C**) and (**D**), the same models are rotated by 90° according to the vertical axis.

7. CBMs/structural motifs annotation and phylogenetic analysis of GH19

In order to obtain more insight into the distribution of accessory domains, the following list of CBMs and other peptidoglycan/chitin binding modules were annotated in the database: CBM18, CBM5/12 and CBM13, LysM, PG_binding_1, PGRP and SH3_3. Their presence in association with catalytic domains was plotted on the networks of homologous families, previously obtained. For this step, CBM18, CBM5/12, CBM13, LysM and PGRP profile HMMs were built using HMMER from the multiple sequence alignments available in the SMART database [42] with accession codes SM00270 (CBM18), SM00495 (CBM5/12), SM00458 (CBM13), SM00257 (LysM) and SM00701 (PGRP). The profile HMMs available in Pfam with the accession codes PF01471 and PF08239 were used for PG_binding_1 and SH3_3, respectively. The annotations were performed by scanning the GH19ED database with these profile HMMs; scanning parameters were more permissive than the ones used for the catalytic domain, as CBMs are shorter and contain more repetitive motifs, so more sensitivity is required (maximal E-value of 10^{-5}, minimum length of 20 residues and bias ratio lower than 1).

The work in [141] reported the presence of a 3-helix peptidoglycan binding bundle (named PBM in this work) in the reference endolysin from bacteriophage SPN1S, which is the only functionally characterized structural motif in GH19 endolysin catalytic domain; in order to annotate this motif in the database, a sequence-based alignment was obtained with MAFFT "L-INS-i" strategy [290] from up to 600 close homologs searched in Uniprot non-redundant Uniref50 database, by using as query the reference endolysin and a restrictive E-value threshold of 10^{-20}. The alignment was manually cut with respect to the length of the reference PBM and used to generate the profile HMM used for annotations in GH19ED, with the same criteria employed above for CBMs.

The standard positions corresponding to the six chitinases reference loops were annotated in the GH19ED by looking to recent GH19 literature with respect to the corresponding motifs present in the reference "loopful" plant chitinase from rye seed (PDB accession 4j0l) and absent from the structurally aligned "loopless" plant chitinase from *Gemmabryum coronatum* (PDB accession 3wh1), as shown in **Fig. 2A**. The minimum length allowed for a loop was 4 residues shorter with respect to the loop length of the reference.

A coarse-grained phylogeny was built from catalytic domain centroids extracted from all homologous families in the defined superfamilies. Centroids were defined by CD-HIT [291] heuristic clustering algorithm at 40% identity cut-off (and word size 2).

A starting approximate centroids alignment was built with the E-ins-I algorithm of MAFFT. A Bio-Neighbour Joining [241] starting tree was generated from this alignment through Phylogeny.fr web service (http://www.phylogeny.fr/). These results were refined in a Bayesian analysis by Bali-Phy 3.4 [242]. Six independent Monte Carlo Markov chain analyses were run and stopped after 200,000 cycles, when sampled parameters resulted in convergence and good mixing (http://www.bali-phy.org/README.html#mixing_and_convergence). In order to eliminate the background noise at the beginning of the run, the first 50% of samples were discarded. Each analysis was performed at default parameters priors with an LG empirical substitution rate matrix and an rs07 [244] insertion/deletion model. The resulting unrooted tree is the majority consensus from all the samples collected during the runs. The position of the root was obtained by considering the splitting between superfamily networks (identified as described in paragraph 3), if supported in the obtained phylogeny.

Author Contributions

M. O. performed the experiments, analysed data and wrote the manuscript; P.C.F.B. provided scripts and resources to set up and analyse the database, performed network degree and size distribution analyses and contributed to writing the manuscript; M. L. supervised the project and contributed to writing the manuscript; J. P. planned the experiments, supervised the project and contributed to writing the manuscript.

Acknowledgements

M.O. acknowledges a PhD fellowship by the University of Milano-Bicocca, P.C.F.B and J.P. acknowledge funding by Bundesministerium für Bildung und Forschung (grant 031B0571A) and by Deutsche Forschungsgemeinschaft (grant EXC2075).

SUPPLEMENTARY MATERIAL

Table S1. List of biochemically characterized GH19 seed sequences screened from literature and used for BLAST searches to initialize the GH19ED database. Superfamily assignments are based on **Fig. 1**. Homologous family assignments and numeral identifier are based on **Fig. 2**. hfam ID = homologous family numeral identifier.

Uniprot Accession	PDB Accession	Source	N° (CBM)	Superfamily (hfam ID)	Activity	Property	References
P29022	4mck	*Zea mais* (Plants)	1 (18)	CHIT (2A)	Chitinase, allergen	Antifungal	[217, 292, 293]
O64203		*Mycobacterium* phage D29 (Virus)	[a]	ELYS (8)	Lysozyme		[193]
A9ZSX9	3wh1	*Gemmabryum coronatum* (Plants)		CHIT (2B)	Chitinase		[111, 122, 123, 155, 207]
Q9WXI9		*Aeromonas sp.* 10S24 (Gammaproteobacteria)	2 (5/12)	CHIT (7)	Chitinase		[128, 274]
R0MMH7		*Nosema bombycis* (Fungus)	[b]	CHIT (14)	Chitinase		[223]
O50152	1wvu	*Streptomyces griseus* HUT 6037 (Actinobacteria)	1 (5/12)	CHIT (5)	Chitinase	Antifungal	[124, 261, 294-296]
Q9SQF7	2z37	*Brassica juncea* (Plants)	2 (18)	CHIT (1)	Chitinase, allergen		[275, 297]
Q9FRV0	4j0l	*Secale cereale* (Plants)		CHIT (1)	Chitinase	Antifungal	[139, 201, 203, 298-300]
Q5J1K1		*Streptomyces sp.* MG3 (Actinobacteria)	1 (5/12)	CHIT (5)	Chitinase	Antifungal	[301]
Q25BT4		*Vibrio proteolyticus* (Gammaproteobacteria)	2 (5/12)	CHIT (6)	Chitinase		[129]
A4C3H5		*Pseudoalteromonas tunicata* (Gammaproteobacteria)	1 (5/12)	CHIT (6)	Chitinase	Antifungal	[130]
Q43752		*Citrus sinensis* (Plants)		CHIT (1)	Chitinase		[302]
Q9XFW7		*Beta vulgaris* (Plants)	1 (18)	CHIT (2A)	Chitinase	Antifungal	[303-305]
P25765		*Oryza sativa* (Plants)	1 (18)	CHIT (1)	Chitinase	Antifungal	[266]
Q9SAY3		*Oryza sativa* (Plants)	1 (18)	CHIT (1)	Chitinase		[306]
Q9FEW1		*Nicotiana tabacum* (Plants)	1 (18)	CHIT (1)	Chitinase, lysozyme		[117, 202, 307]
Q9AXR8		*Secale cereale* (Plants)		CHIT (1)		Ice growth inhibition	[308]
P85084	3cql	*Carica papaya* (Plants)		CHIT (1)	Chitinase		[213, 309]
P23951	2baa	*Hordeum vulgare* (Plants)		CHIT (1)	Chitinase	Antifungal	[110, 144, 215, 218, 219, 310-313]
Q7DNA1	31vr	*Oryza sativa* (Plants)	1 (18)	CHIT (1)	Chitinase	Antifungal	[220, 265, 314, 315]
Q8MD06		*Leucaena leucocephala* (Plants)	1 (18)	CHIT (1)	Chitinase	Antifungal	[316]
Q42995		*Oryza sativa* (Plants)	1 (18)	CHIT (1)	Chitinase	Antibacterial (expression strain)	[317]
Q9AXR9		*Secale cereale* (Plants)	1 (18)	CHIT (1)		Ice growth inhibition	[308]
P42820		*Beta vulgaris* (Plants)	1 (18)	CHIT (2A)	Chitinase		[304, 318]
Q4KHC5		*Pseudomonas fluorescens* Pf-5 (prophage in bacteria)		ELYS (1)	Lysozyme		[319]
A7UC81		*Oryza sativa* (Plants)	1 (18)	CHIT (1)	Chitinase		[320]

P11218		Urtica dioica (Plants)	2 (18)	CHIT (4)	Allergen	Antifungal, insecticidal	[132, 272, 321, 322]
V5TEI0		*Dionea muscipula* (Plants)	1 (18)	CHIT (1)	Chitinase	Insect digestion	[323]
Q96408		*Daucus carota* (Plants)	1 (18)	CHIT (2A)	Chitinase		[324]
G9B4E2		*Picea engelmannii* x *Picea glauca* (Plants)	1 (18)	CHIT (1)	Chitinase		[325]
Q9RHU5		*Streptomyces thermoviolaceus* OPC-520 (Actinobacteria)	1 (13)	CHIT (5)	Chitinase	Antifungal	[326]
G3BM11		*Salmonella enterica* phage PVPSE1 (Phage)	[c]	ELYS (3)	Lysozyme	Antibacterial	[136]
Q59I46		*Bacillus circulans* (Firmicutes)	2 (5/12)	CHIT (5)	Chitinase		[327-329]
Q9Z9M4	2cjl	*Streptomyces coelicolor* A3 (Actinobacteria)		CHIT (5)	Chitinase		[125, 330, 331]
G9B4E3		*Picea engelmannii* x *Picea glauca* (Plants)	1 (18)	CHIT (1)	Chitinase		[325]
Q9RHU4		*Streptomyces thermoviolaceus* OPC-520 (Actinobacteria)		CHIT (5)	Chitinase	Antifungal	[326]
B1B6T0		*Bromus inermis* (Plants)	1 (18)	CHIT (1)	Chitinase	Plant cold-response	[332]
Q9FUH3	4tx7	*Vigna unguiculata* (Plants)	1 (18)	CHIT (1)	Chitinase	Antifungal	[214]
O04138		*Oryza sativa* (Plants)	1 (18)	CHIT (2A)	Chitinase	Antifungal	[221]
Q42428		*Castanea sativa* (Plants)	1 (18)	CHIT (1)	Chitinase	Antifungal	[333, 334]
O81934	1dxj	*Canavalia ensiformis* (Plants)		CHIT (1)	Chitinase		[335, 336]
P29023		*Zea mais* (Plants)	1 (18)	CHIT (2A)	Chitinase	Antifungal	[337]
G9I9L2		*Pseudomonas* phage OBP (Phage)	[c]	ELYS (5)	Lysozyme	Antibacterial	[136]
B3VFX0		*Limonium bicolor* (Plants)	1 (18)	CHIT (1)	Chitinase	Antifungal	[338]
F8WSX8		*Chitiniphilus shinanonensis* (Betaproteobacteria)	2 (5/12)	CHIT (5)	Chitinase	Antifungal	[339]
O24530		*Vitis vinifera* (Plants)	1 (18)	CHIT (2A)	Chitinase		[340]
Q9FRV1		*Secale cereale* (Plants)	1 (18)	CHIT (1)	Chitinase	Antifungal	[203, 298, 341, 342]
P17513		*Nicotiana tabacum* (Plants)		CHIT (1)	Chitinase		[117, 202, 343]
Q9ZTT8		*Gossypium hirsutum* (Plants)	1 (18)	CHIT (1)	Chitinase		[344]
O23804		*Chenopodium amaranticolor* (Plants)	1 (18)	CHIT (2A)	Chitinase	Antibacterial (expression strain)	[204]
Q9LBM0		*Burkholderia gladioli* (Betaproteobacteria)	1 (5/12)	CHIT (5)	Chitinase		[127]
Q6WSR8	3hbe	*Picea abies* (Plants)	1 (18)	CHIT (2A)			[121, 345]
P19171		*Arabidopsis thaliana* (Plants)	1 (18)	CHIT (1)		Pathogen resistance	[346, 347]
Q43184		*Solanum tuberosum* (Plants)		CHIT (1)	Chitinase		[348]
O24531		*Vitis vinifera* (Plants)	1 (18)	CHIT (2A)	Chitinase		[340]
Q949H3	4mst	*Hevea brasiliensis* (Plants)	1 (18)	CHIT (1)	Allergen	Antifungal	[212, 349]
B3XZQ2		*Streptomyces cyaneus* SP-27 (Actinobacteria)	1 (5/12)	CHIT (5)	Chitinase	Antifungal (protoplast formation)	[350, 351]
P24626		*Oryza sativa* (Plants)	1 (18)	CHIT (1)	Chitinase	Antifungal	[216, 221]

P08252		*Nicotiana sylvestris* (Plants)	1 (18)	CHIT (1)	Chitinase, lysozyme	Antifungal	[202, 205, 262, 307, 352-354]
P17514		*Nicotiana tabacum* (Plants)		CHIT (1)	Chitinase, lysozyme		[117, 202, 343]
Q2HJJ5		*Musa paradisica* (Plants)	1 (18)	CHIT (1)	Chitinase	Antifungal	[225]
A0A516Z9V1		*Pseudomonas sp.* Ef1 (prophage in bacteria)		ELYS (1)	Lysozyme		[d]
Q42878		*Solanum lycopersicum* (Plants)		CHIT (1)	Chitinase		[355]
G9B4E8		*Pinus contorta* (Plants)	1 (18)	CHIT (1)	Chitinase		[325]
Q8GI53		*Nocardiopsis prasina* (Actinobacteria)	1 (5/12)	CHIT (5)	Chitinase	Antifungal	[356]
A0A516Z9W0		*Pseudomonas sp.* Ef1 (prophage in bacteria)		ELYS (1)	Lysozyme	Antibacterial	[d]
Q5NTA4	5h7t	*Cryptomeria japonica* (Plants)	1 (18)	CHIT (2A)	Allergen, Chitinase	Antifungal	[263, 357, 358]
B5L6N2		*Crocus sativus* (Plants)	1 (18)	CHIT (1)	Chitinase	Antifungal	[226]
Q207U1		*Momordica charantia* (Plants)	1 (18)	CHIT (1)		Antifungal (transgenic)	[267]
Q9S566		*Pseudomonas aeruginosa* PAO1 (prophagic Virus)		ELYS (1)	Lysozyme		[359]
H2D0G4	4ok7	*Salmonella phage* SPN1S (Phage)		ELYS (2)	Lysozyme	Antibacterial	[113, 141]
A0A0M4F9K9		*Acinetobacter* phage vb_AbaP_CEB1 (Phage)		ELYS (6)	Lysozyme	Antibacterial	[112]
F1BCP4		*Acinetobacter* phage ΦAB2 (Phage)		ELYS (6)	Lysozyme	Antibacterial	[133, 134]
B2ZY61		*Ralstonia* phage ΦRSL1 (Phage)		ELYS (4)		Cell shape modification (expression strain)	[138]
A0A7I3		*Microcystis aeruginosa* phage (Phage)		ELYS (7)	Chitinase, lysozyme		[137]

[a]This sequence contains an N-terminal amidase domain and a C-terminal domain for host specific membrane binding.

[b]This sequence presents a N-terminal domain with unknown function, potentially involved in chitin binding. [c]These sequences possess an N-terminal putative peptidoglycan binding module.

[d]Biochemical data on these two endolysins has not been published yet. CBM = Carbohydrate binding module.

Table S2. List of GH19 superfamilies and homologous families defined in this study (**Figs. 1** and **2**), their respective number of sequences and proteins (99% identity clustering of sequences), and the average number of residues in the catalytic domain ± standard deviation.
h-fam ID = homologous family numeral identifier in **Fig. 2**.

Superfamily	h-fam ID	Literature-based classification[b]	N° sequences	N° *proteins*	Av. n° residues (±SD)
Chitinases (CHIT, 8554 sequences)	1	class I and II, or plant "loopfull"	1356	1087	227 ± 30
	2A - 2B	class IV, or plant "loopless"	1253	971	195 ± 17
	3		402	323	236 ± 24
	4	class I and II	31	27	226 ± 14
	5	class IV, or bacteria "loopless"	1854	1311	202 ± 9
	6 - 7	cluster IV	2652	969	243 ± 19
	8 to 12		317	253	212 ± 21
	13 -14		33	31	146 ± 7
	15		18	13	161 ± 4
	16 -17		114	94	187 ± 18
	Unclassified[a]		524	482	
Endolysins (ELYS, 10967 sequences)	1	cluster III	3466	2490	181 ± 17
	2	cluster III	2780	1876	196 ± 11
	3	cluster III	10	10	179 ± 1
	4	cluster III	1	1	200
	5	cluster III	8	8	192 ± 14
	6	cluster III	24	23	163 ± 3
	7	cluster III	1	1	165
	8	cluster III	246	166	170 ± 4
	9 to 34	cluster III	1813	1418	178 ± 17
	Unclassified[a]		2620	2232	

[a]Sequences contained in smaller groups were not classified but were still inserted in the database.

[b]The classification system from literature is based on [118, 119, 126], with possible inconsistencies due to independent sources and the higher sequence sampling contained in this study.

Table S3. Most conserved sites according to Rate4site analysis (see *Experimental Procedures* section) in CHIT superfamily. Standard position numbering is according to the chitinase from rye seed (PDB accession 4j0l). Information is provided about the % frequency of amino acids (if higher than 1%) at each site, and the respective function (if known from [139]). Standard positions corresponding to conserved sites in ELYS superfamily (**Tab. S4**) are highlighted in bold. Standard positions specific for CHITs are marked in red.

Standard position	> 90% non-gapped sequences[a]	Conserved residues			Function[b]
6		V 38%	I 15%	L 4.4%	
11		F 58%	W 3.6%	Y 1.9%	
14		M 31%	I 18%	L 13%	
18		R 45%	A 13%	K 2.4%	
23		C 30%	A 4.4%		
28		F 58%	E 7.1%	A 2.9%	
29		Y 86%	W 4.4%		
30		T 73%	S 11%	D 4.5%	
31		Y 77%	R 13%	F 1.1%	
34		F 67%	L 25%		
37		A 90%			
44	X	F 80%	V 8.1%	L 5.6%	
54	X	K 63%	R 25%	M 3.8%	
55	X	R 42%	K 35%	K 15%	
56	**X**	**E 68%**	**S 23%**	**T 5.2%**	
58	**X**	**A 80%**	**V 11%**	**I 4.2%**	
59	**X**	**A 63%**	**T 32%**		
60	**X**	**F 61%**	**M 19%**	**A 16%**	
61	X	F 53%	L 43%		
62	**X**	**A 82%**	**G 12%**	**T 1.8%**	
63	**X**	**H 52%**	**N 25%**	**Q 19%**	
64	X	F 35%	V 34%	T 15%	
66	**X**	**H 54%**	**Q 34%**	**S 3.6%**	**Substrate binding (+1)**
67	**X**	**E 91%**	**K 4.7%**		**Catalytic proton donor and substrate binding (-1)**
68	**X**	**T 91%**	**S 5.4%**		
84	X	L 69%	Y 14%	F 12%	
89	**X**	**E 94%**			**Catalytic base and substrate binding (-1)**
96	**X**	**Y 90%**	**K 1.7%**	**M 1%**	**Substrate binding (-1)**
97	**X**	**C 93%**	**V 1.5%**		
105	**X**	**C 91%**	**P 3.8%**	**G 1.2%**	
111	X	Y 99%			
113	**X**	**G 99%**			
114	**X**	**R 98%**	**K 1.4%**		
115	**X**	**G 99%**			
116	X	P 59%	A 38%		
118	**X**	**Q 91%**	**P 4.8%**	**M 3.2%**	**Substrate binding (+1)**
120	**X**	**S 84%**	**T 9.6%**	**Y 4.3%**	**Catalytic water coordination and substrate binding (-2)**
122	X	N 82%	H 15%		
124	**X**	**N 99%**			**Substrate binding (-2)**
125	**X**	**Y 99%**			
128	X	A 54%	F 36%	C 5.6%	
129	X	G 62%	S 37%		
136	X	L 83%	G 15%		

137	X	L 89%	I 5.6%		
140	**X**	**P 100%**			
143	**X**	**V 91%**	**I 4.2%**	**L 3.8%**	
150	X	A 38%	N 33%	S 26%	
151	**X**	**F 41%**	**L 32%**	**W 23%**	
153	X	T 52	S 37%	A 8.3%	
154	**X**	**A 85%**	**G 13%**	**S 1.3%**	
156	X	W 79%	F 18%	L 1.4%	
158	**X**	**W 66%**	**F 30%**	**Y 2.3%**	
159	X	M 42%	N 19%	L 15%	
165		K 53%	T 16%	S 4.6%	
167		S 45%	T 22%	N 10%	
169	X	H 60%	L 32%	R 4.2%	
171	X	V 66%	A 27%	I 2.5%	
190	**X**	**G 95%**	**N 1.8%**		
191	X	F 84%	Y 13%		
192	**X**	**G 95%**	**A 3.5%**		
194	X	T 81%	I 10%	V 4.4%	
195	**X**	**I 63%**	**T 29%**	**M 4.3%**	
196	X	N 40%	R 35%	Q 16%	
197	X	I 45%	S 19%	A 16%	
198	**X**	**I 88%**	**L 8.2%**	**V 2.1%**	**Substrate binding without side chain (-2)**
199	**X**	**N 92%**	**Y 4.8%**		**Substrate binding (-2)**
200	**X**	**G 94%**	**S 2.3%**	**A 1.8%**	
203	X	E 92%	V 2.9%		Substrate binding (+1)
204	X	C 97%			
215	**X**	**R 92%**	**I 4.3%**		**Substrate binding (+1)**
216	X	I 66%	V 29%	Y 1.1%	
219	X	Y 75%	W 15%	F 7.7%	
222	**X**	**F 39%**	**Y 34%**	**L 15%**	
223	X	A 35	C 31%	T 17%	
228	X	V 69%	I 15%	T 8.9%	
231		G 60%	D 15%	P 7.5%	
236		C 85%			

[a]All CHITs sequences classified in the GH19ED database were considered.

[b]Binding subsites (in parenthesis) are numbered according to the standard nomenclature; cleavage occurs between the sugar units bound at subsites -1 and +1 [140].

Table S4. Most conserved sites according to Rate4site analysis (see *Experimental Procedures* section) in ELYS superfamily. Standard position numbering is according to the endolysin from bacteriophage SPN1S of Salmonella typhimurium (PDB accession 4ok7). Information is provided about the % frequency of amino acids (if higher than 1%) at each site, and the respective function (if known from [141]). Standard positions corresponding to conserved sites in CHIT superfamily (**Tab. S3**) are highlighted in bold. Standard positions specific for ELYSs are marked in red.

Standard position	> 90% non gapped sequences[a]	Conserved residues			Function
33	**X**	**I 80**	**L 3.3%**	**C 3.2%**	
38	**X**	**R 60%**	**D 22%**	**W 4.5%**	
40	**X**	**A 90%**	**S 2.1%**	**C 1.8%**	
41	**X**	**A 36%**	**M 34%**	**Y 9.3%**	
42	**X**	**F 82%**	**W 4.5%**	**M 3.7%**	
44	**X**	**A 91%**	**G 4.3%**	**S 3.4%**	
45	**X**	**Q 88%**	**T 9.3%**		
48	**X**	**H 93%**	**V 2.3%**		
49	**X**	**E 99%**			**Catalytic proton donor**
50	**X**	**S 86%**	**T 11%**	**C 1.4%**	
53	X	L 44%	F 43%	M 7.3%	
58	**X**	**E 99%**			**Catalytic base**
61		N 53%	G 15%	S 8.9%	
63		S 47%	T 17%	A 16%	
102	X	A 62%	Q 3.6%	L 3.5%	
106	**X**	**Y 95%**	**F 2.3%**		
109	**X**	**R 80%**	**E 5.4%**	**A 4.2%**	
110	X	L 51%	M 25%	N 16%	
111	X	G 94%	V 1.6%		
112		N 83%	D 10%		
116	X	G 83%	T 5.8%	A 2%	
117	X	D 93%	E 2.6%		
118	**X**	**G 97%**	**Y 1.4%**		
122	X	R 79%	K 9.3%	L 3.8%	
123	**X**	**G 99%**			
124	**X**	**R 95%**	**G 1.2%**	**A 1.2%**	
125	**X**	**G 96%**	**T 2.5%**		
126	X	L 75%	P 7.7%	Y 4.3%	
127	X	I 63%	L 14%	V 8.9%	
128	**X**	**Q 92%**	**M 5.5%**	**G 1%**	
130	**X**	**T 99%**			**Catalytic water coordination**
131	X	G 85%	F 5.3%	W 4.3%	
134	**X**	**N 96%**			
135	**X**	**Y 97%**	**F 1.5%**		
150	**X**	**P 95%**	**G 1.2%**		
153	**X**	**L 74%**	**A 13%**	**V 7.8%**	
159	X	A 82%	S 6.2%	G 5.5%	
163	**X**	**A 83%**	**S 4.2%**	**E 1.9%**	
167	**X**	**W 65%**	**F 14%**	**Y 10%**	
172	X	L 60%	C 29%	I 4.5%	
173	**X**	**L 54%**	**Y 23%**	**F 4.5%**	
176		R 19%	A 11%	S 11%	
183	**X**	**T 81%**	**R 9.6%**	**S 3.6%**	
186	**X**	**I 85%**	**V 12%**		
187	**X**	**N 97%**			
188	**X**	**G 89%**	**L 3.3%**	**P 1.5%**	

189		G 83%	A 2.1%	R 1.1%
191		N 75%	T 3.4%	E 2.7%
192		G 85%		
196	**X**	**R 89%**		
203		A 74%	I 3.6%	C 2.1%

[a]All ELYSs sequences classified in the GH19ED database were considered.

Table S5. Exemplary sequence entries from the GH19ED with degrees n > 300 in hub regions formed at 95% sequence identity for the conserved core, listed with their corresponding annotation, taxonomic name of the source organism and NCBI accession (compare with **Fig. S8**).

Degree	Annotation	Source organism	NCBI accession
301	glycoside hydrolase family 19 protein	*Pseudomonas aeruginosa*	WP_116825151.1
308	glycoside hydrolase family 19 protein	*Pseudomonas aeruginosa*	WP_123789282.1
355	carbohydrate-binding protein	*Vibrio atlanticus*	WP_065678462.1
364	carbohydrate-binding protein	*Vibrio owensii*	WP_122068128.1
376	carbohydrate-binding protein	*Vibrio splendidus*	WP_108214678.1
376	carbohydrate-binding protein	*Vibrio splendidus*	WP_102462326.1
376	carbohydrate-binding protein	*Vibrio splendidus*	WP_108123460.1
395	carbohydrate-binding protein	*Vibrio chagasii*	WP_128161859.1
395	carbohydrate-binding protein	*Vibrio splendidus*	WP_116870071.1

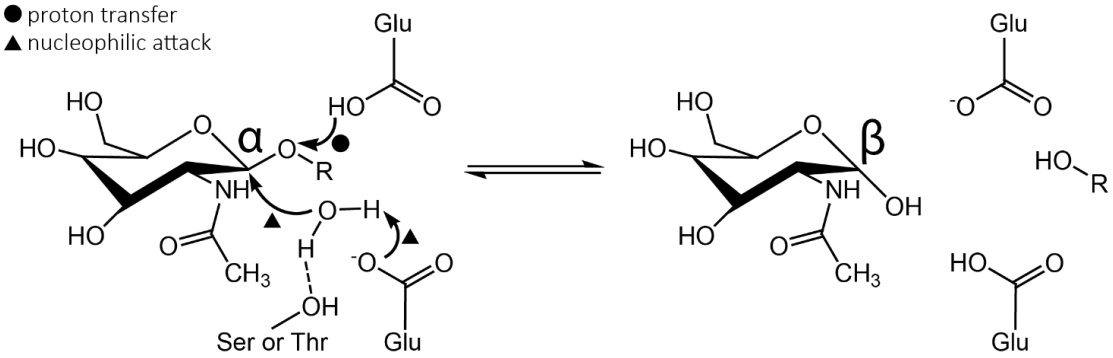


Figure S1. The single displacement hydrolysis mechanism of GH19 [110]. One acidic, one basic glutamate and a serine (or threonine) for water placement are generally required in the active site and the hydrolysis product has inversion of the anomeric configuration.

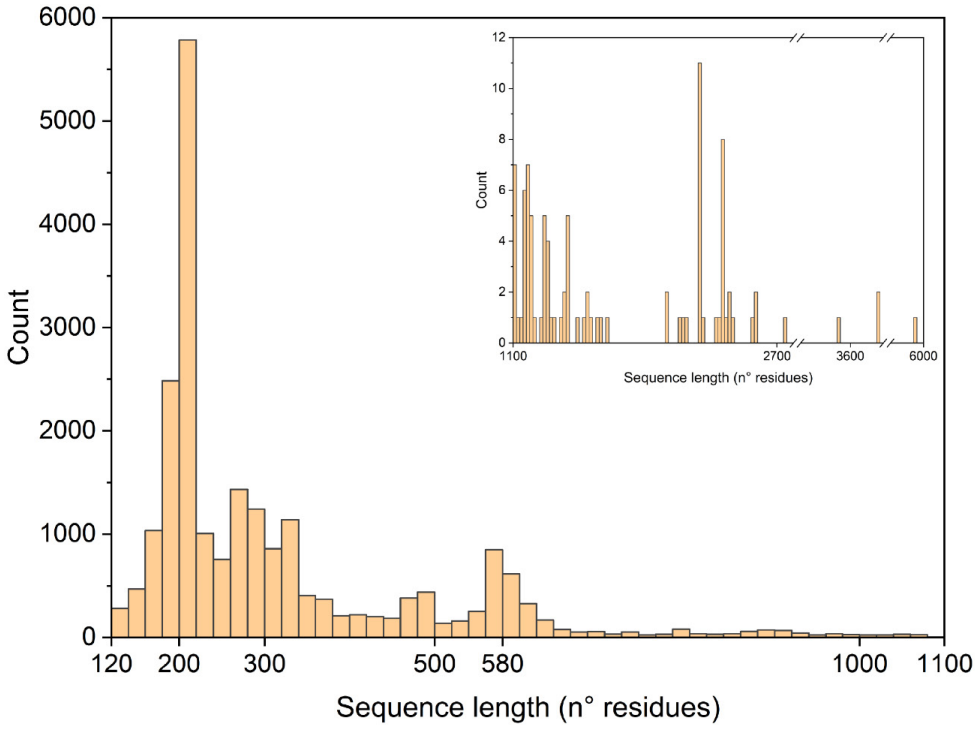


Figure S2. Length distribution histogram of sequence entries in the GH19ED database, with a bin size of 20 residues. The two main peaks are around 200 and 580 residues. Only few sequences are longer than 1100 residues up to 6000.

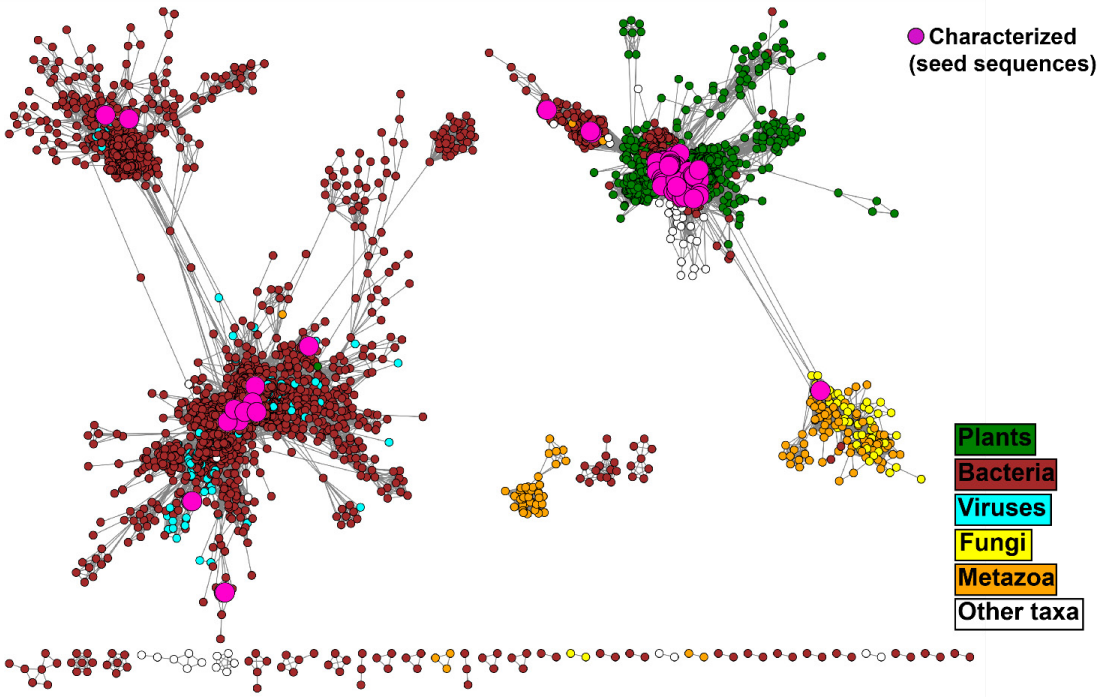


Figure S3. Protein sequence networks of all GH19 representative domains (5229 sequence node centroids obtained from clustering at 90% identity) connected by edges with an identity cut-off of 40%. The two bigger clusters contain seed sequences of characterized endolysins (2738 sequence nodes on the left) and chitinases (2329 sequence nodes on the right). The prefuse force-directed OpenCL layout was used. The domains were extracted from Pfam GH19 HMM profile (PF00182) scanning of the sequences collected through BLAST searches in which the seed sequences reported in **Tab. S1** were used as queries. Nodes are colored according to their annotated taxonomic source. In **Fig. 1** only the two main clusters are visualized.

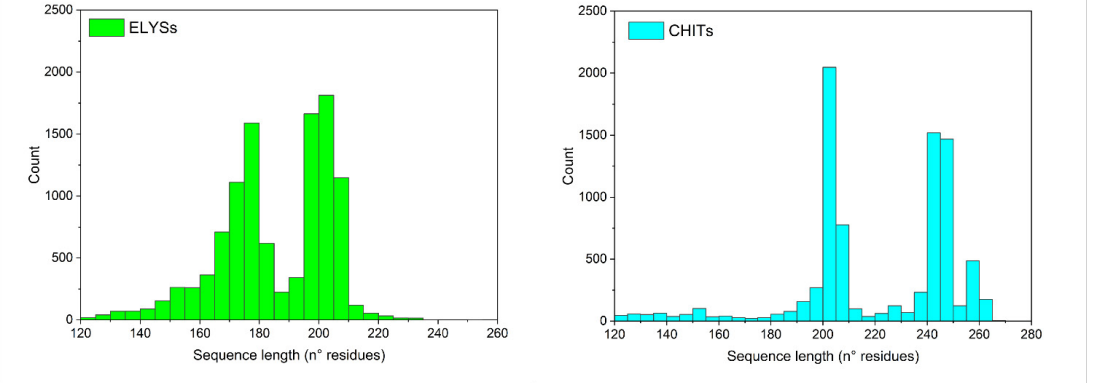


Figure S4. Length distribution histogram of ELYS and CHIT domains in the GH19ED database, with a bin size of 5 residues. The two main peaks are around 175 and 200 for ELYSs, 200 and 245 for CHITs.

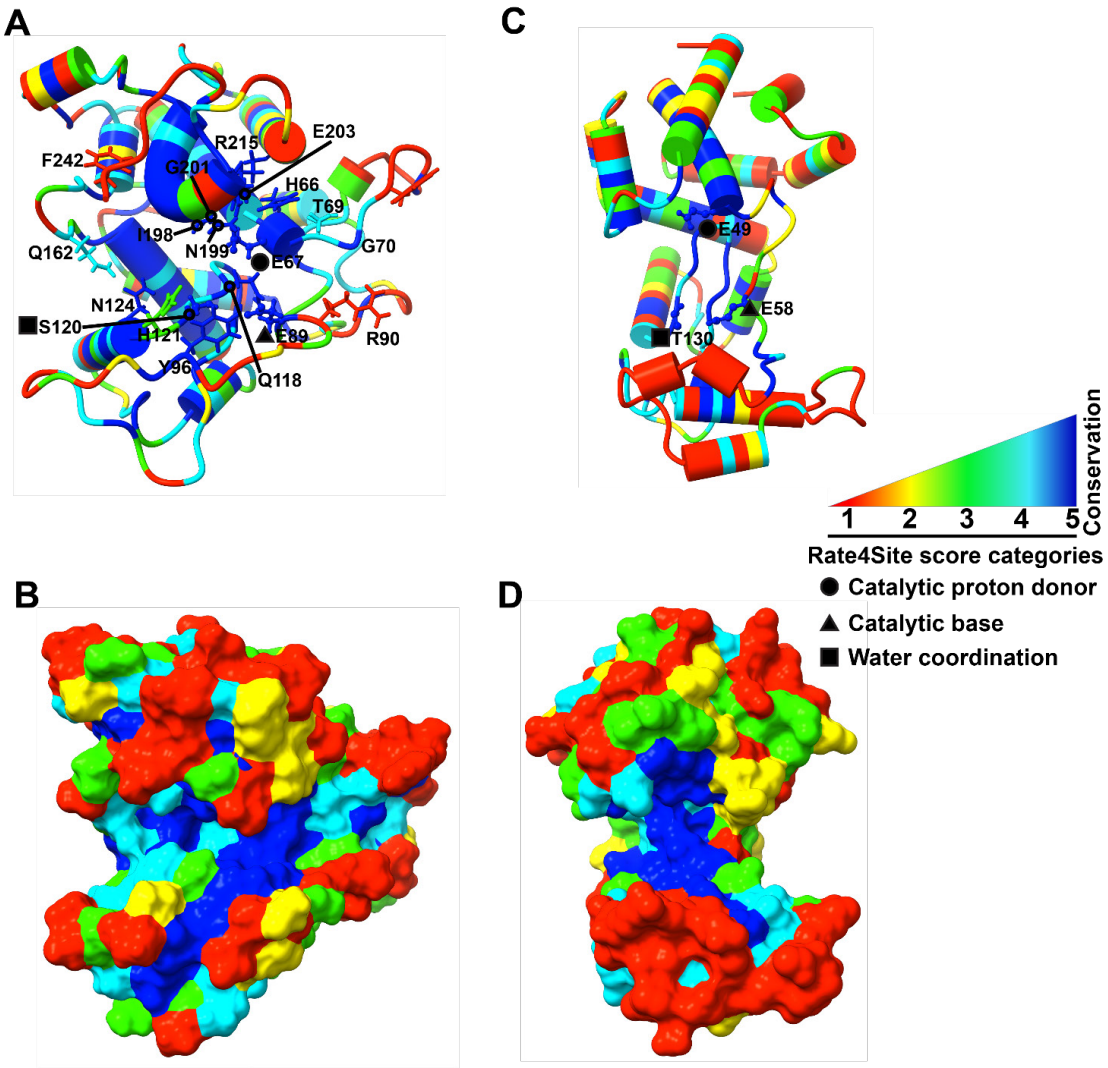


Figure S5. Rate4Site categories of conservation (see *Experimental Procedures* section) are visualized onto models of CHIT reference (**A**-**B**, PDB accession 4j0l) and ELYS reference structure (**C**-**D**, PDB accession 4ok7). (**A**) and (**C**) models are visualized as cartoon with α-helices shown as cylinders, substrate binding residues as sticks (except glycine), and catalytic residues as balls and sticks. (**B**) and (**D**) are the same models shown in **A** and **C**, represented as solvent accessible surface areas.

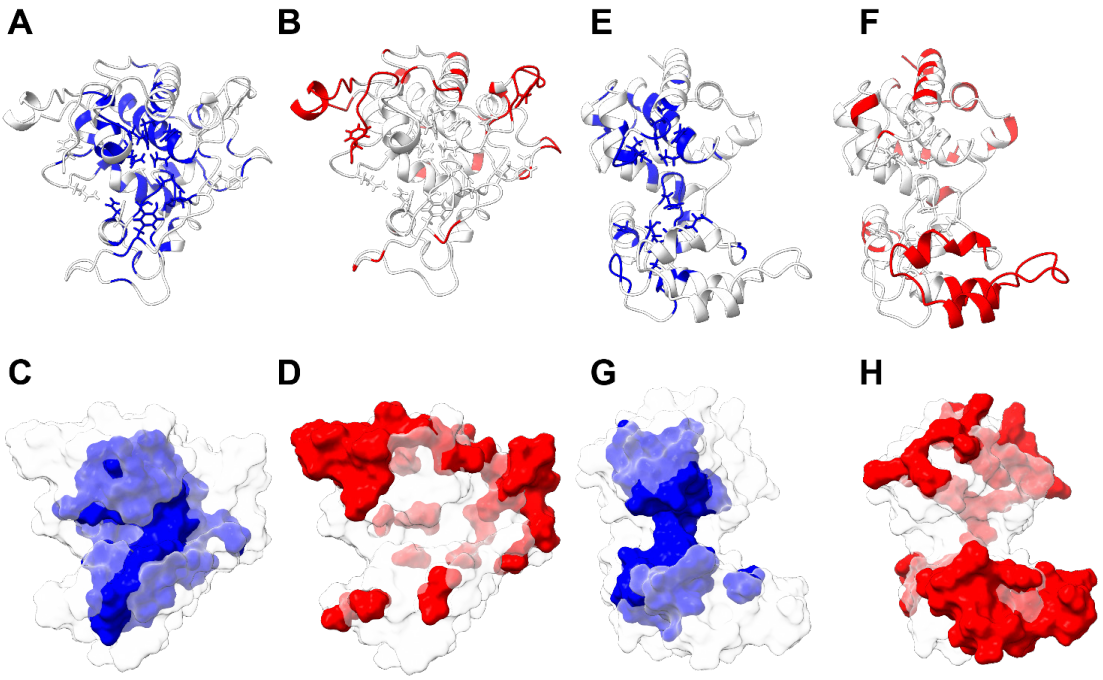


Figure S6. Rate4Site conservation categories 1 (least conserved) and 5 (most conserved), as declared in *Experimental Procedures* section, are visualized with two different colors (red for 1 and blue for 5) plotted onto 3D models of rye seed CHIT reference (**A**-**C** for category 5 and **B-D** for category 1, PDB accession 4jol) and ELYS reference from bacteriophage SPN1S (**E**-**G** for category 5 and **F**-**H** for category 1, PDB accession 4ok7). (**A**-**B**) The CHIT reference model is visualized in cartoon, with substrate binding residues labelled in **Fig. S5A** as sticks (except for glycine). (**E**-**F**) The ELYS reference model is visualized in cartoon with residues in sticks if corresponding to CHIT substrate binding residues, according to **Tab. 1**. (**C**-**D-G-H**) above models **A**-**B**-**E**-**F** are shown as solvent accessible surface areas.

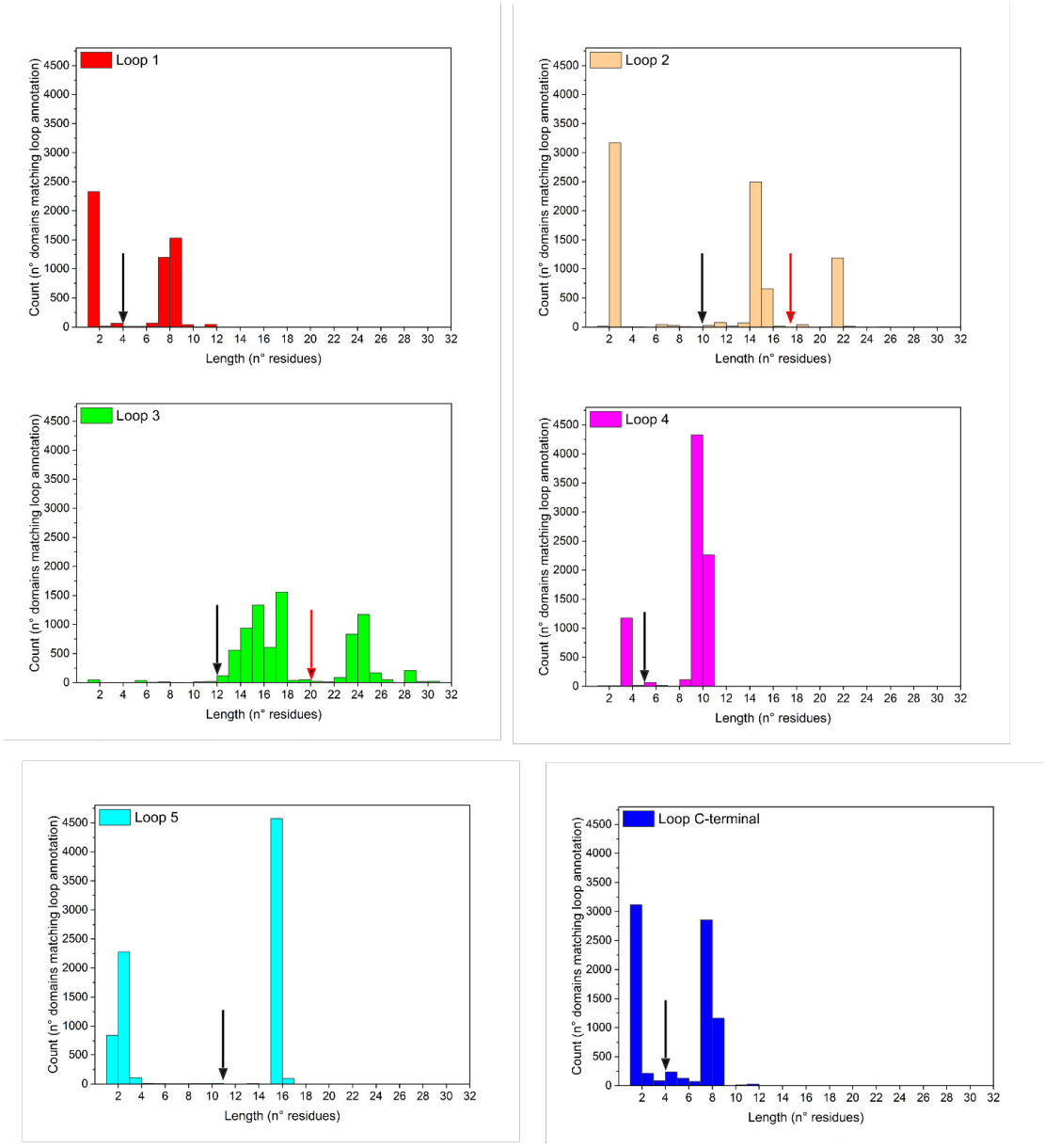


Figure S7. Length distribution of CHIT loop motifs. The black arrow indicates the minimum length threshold used to define the presence of a loop, as specified in the *Experimental Procedures* section. The red arrow indicates the threshold used to separate the two modes of length observed for loops 2 and 3.

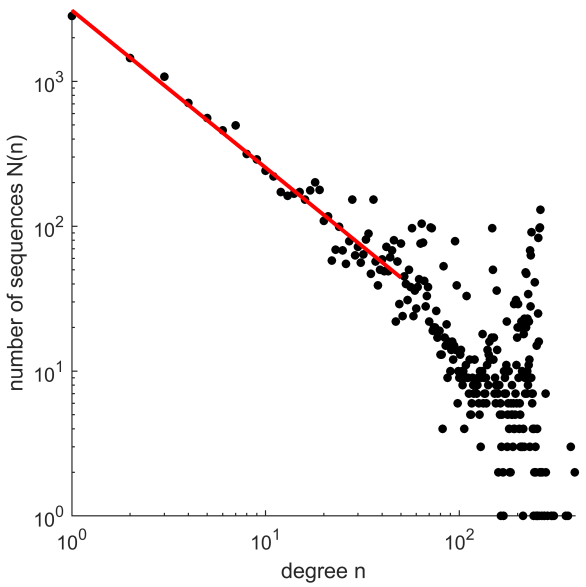


Figure S8. The degree distribution *N(n)* for the conserved region of sequence entries from the GH19ED was approximated by a power-law for degrees ≤ 50 (red line) at a threshold of 95% sequence identity, yielding a scaling exponent of γ = 1.1.

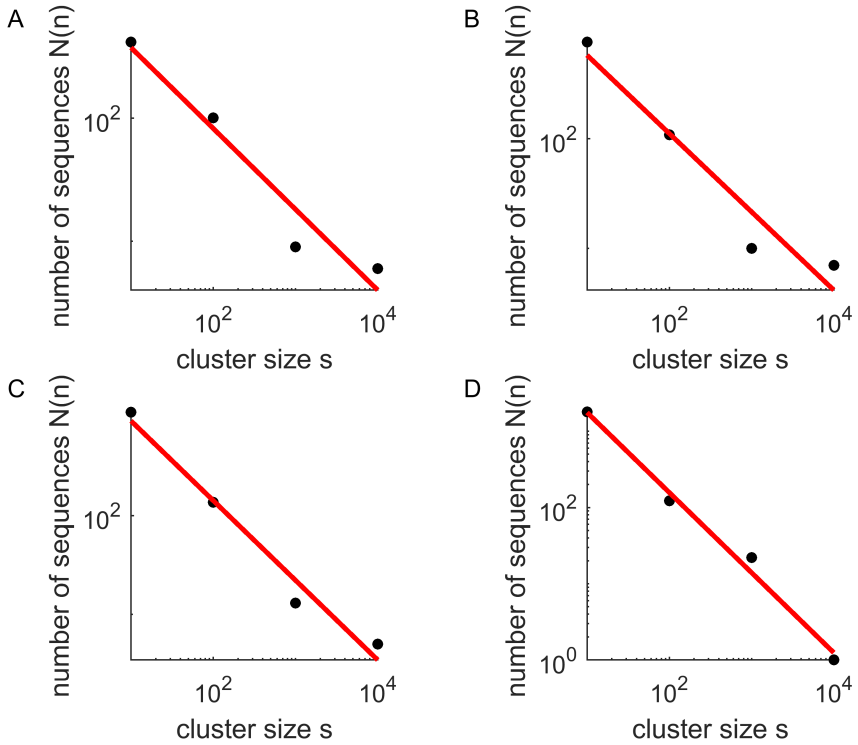


Figure S9. The histograms of the cluster size distributions *N(s)* for the conserved region of sequence entries from the GH19ED at thresholds of 60% (**A**), 70% (**B**), 80% (**C**), and 90% (**D**) sequence identity. The distributions were approximated by a power law yielding exponents τ_h of 0.7, 0.7, 0.8 and 1.1, respectively (compare with **Fig. S10**).

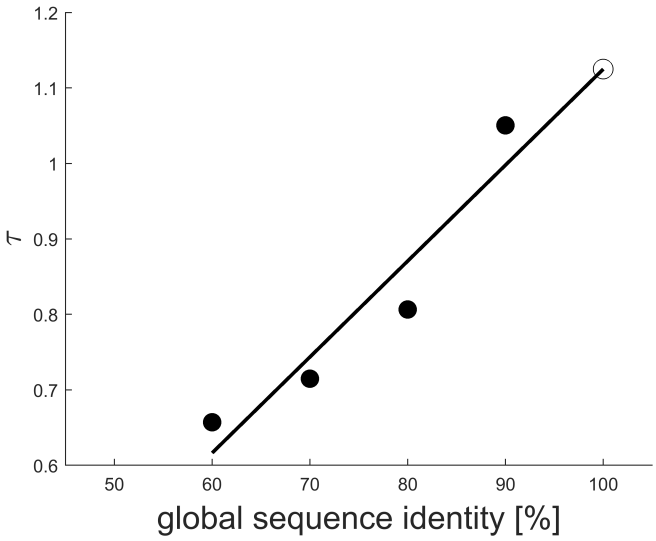


Figure S10. The slopes of the histograms (**Fig. S9**) were used to linearly extrapolate the exponent τ for individual amino acid exchanges at 100% sequence identity.

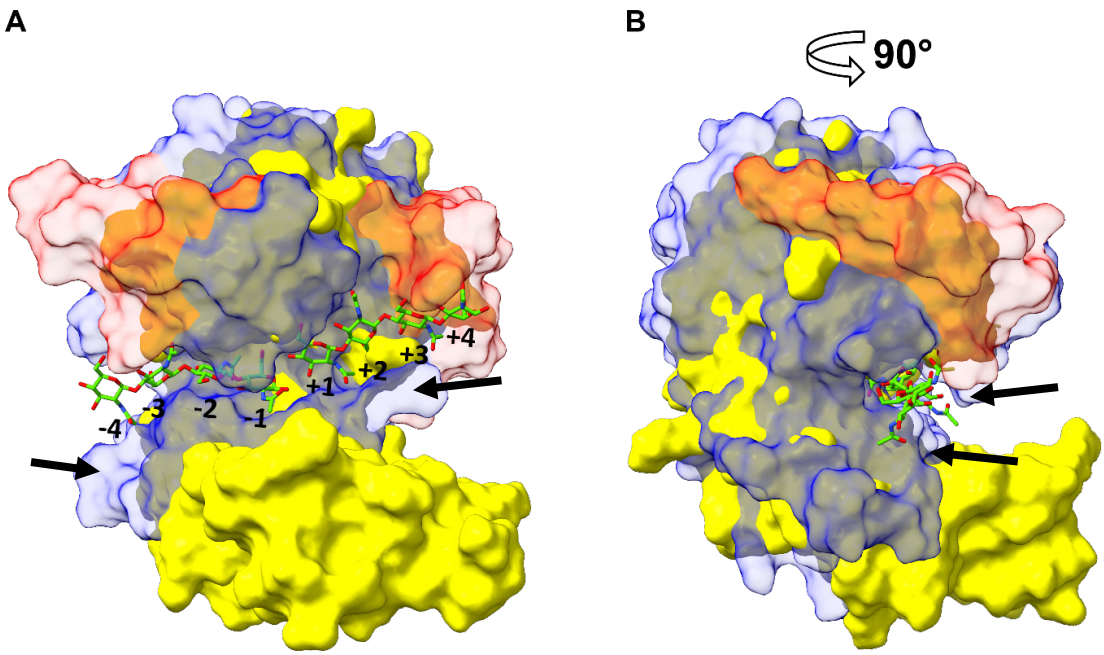


Figure S11. (**A**) The rye seed chitinase model is visualized in blue transparent solvent accessible surface area (loops 1, 2, 5 and C-terminal are colored in red), superposed to the endolysin from bacteriophage SPN1S model, visualized in yellow solvent accessible surface area; two co-crystallized tetrachitooligosaccharides are in the catalytic cleft [139]. (**B**) The same object is rotated by 90° according to the vertical axis. Black arrows highlight the regions in which the cleft of the chitinase model is tighter than the one of the endolysin.

4.References

1. Bornscheuer, U., Huisman, G., Kazlauskas, R. J., Lutz, S., Moore, J. & Robins, K. (2012) Engineering the third wave of biocatalysis, *Nature.* **485**, 185-194.
2. Wohlgemuth, R. (2010) Biocatalysis—key to sustainable industrial chemistry, *Current opinion in Biotechnology.* **21**, 713-724.
3. Saha, B. C. & Demirjian, D. C. (2000) Advances in enzyme development and applied industrial biocatalysis in, ACS Publications.
4. Kamagata, Y. & Tamaki, H. (2005) Cultivation of uncultured fastidious microbes, *Microbes and Environments.* **20**, 85-91.
5. Sekiguchi, Y. (2006) Yet-to-be cultured microorganisms relevant to methane fermentation processes, *Microbes and Environments.* **21**, 1-15.
6. Ekkers, D. M., Cretoiu, M. S., Kielak, A. M. & van Elsas, J. D. (2012) The great screen anomaly—a new frontier in product discovery through functional metagenomics, *Applied microbiology and biotechnology.* **93**, 1005-1020.
7. Festel, G., Detzel, C. & Maas, R. (2012) Industrial biotechnology-Markets and industry structure, *Journal of Commercial Biotechnology.* **18**, 1.
8. Marrs, B., Delagrave, S. & Murphy, D. (1999) Novel approaches for discovering industrial enzymes, *Current opinion in microbiology.* **2**, 241-245.
9. Pennisi, E. (1997) Biotechnology: in industry, extremophiles begin to make their mark in, *Science*. **276**, 705-706.
10. Gomes, J. & Steiner, W. (2004) The biocatalytic potential of extremophiles and extremozymes, *Food technology and Biotechnology.* **42**, 223-225.
11. Elleuche, S., Schroeder, C., Sahm, K. & Antranikian, G. (2014) Extremozymes—biocatalysts with unique properties from extremophilic microorganisms, *Current opinion in biotechnology.* **29**, 116-123.
12. Ferrer, M., Golyshina, O., Beloqui, A. & Golyshin, P. N. (2007) Mining enzymes from extreme environments, *Current opinion in microbiology.* **10**, 207-214.
13. Uchiyama, T., Abe, T., Ikemura, T. & Watanabe, K. (2005) Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes, *Nature biotechnology.* **23**, 88-93.
14. Sjostrom, S. L., Bai, Y., Huang, M., Liu, Z., Nielsen, J., Joensson, H. N. & Svahn, H. A. (2014) High-throughput screening for industrial enzyme production hosts by droplet microfluidics, *Lab on a Chip.* **14**, 806-813.
15. Kaul, P. & Asano, Y. (2012) Strategies for discovery and improvement of enzyme function: state of the art and opportunities, *Microbial biotechnology.* **5**, 18-33.
16. Goodwin, S., McPherson, J. D. & McCombie, W. R. (2016) Coming of age: ten years of next-generation sequencing technologies, *Nature Reviews Genetics.* **17**, 333-351.
17. Li, L.-L., McCorkle, S. R., Monchy, S., Taghavi, S. & van der Lelie, D. (2009) Bioprospecting metagenomes: glycosyl hydrolases for converting biomass, *Biotechnology for biofuels.* **2**, 10.
18. Stewart, J. D. (2006) Genomes as resources for biocatalysis, *Advances in applied microbiology.* **59**, 31-52.
19. Jeon, J. H., Kim, J.-T., Kim, Y. J., Kim, H.-K., Lee, H. S., Kang, S. G., Kim, S.-J. & Lee, J.-H. (2009) Cloning and characterization of a new cold-active lipase from a deep-sea sediment metagenome, *Applied microbiology and biotechnology.* **81**, 865-874.

20. Fernández-Álvaro, E., Kourist, R., Winter, J., Böttcher, D., Liebeton, K., Naumer, C., Eck, J., Leggewie, C., Jaeger, K. E. & Streit, W. (2010) Enantioselective kinetic resolution of phenylalkyl carboxylic acids using metagenome-derived esterases, *Microbial biotechnology.* **3**, 59-64.
21. Knietsch, A., Waschkowitz, T., Bowien, S., Henne, A. & Daniel, R. (2003) Construction and screening of metagenomic libraries derived from enrichment cultures: generation of a gene bank for genes conferring alcohol oxidoreductase activity on Escherichia coli, *Applied Environmental Microbiology.* **69**, 1408-1416.
22. Gabor, E. M., De Vries, E. J. & Janssen, D. B. (2004) Construction, characterization, and use of small-insert gene banks of DNA isolated from soil and enrichment cultures for the recovery of novel amidases, *Environmental Microbiology.* **6**, 948-958.
23. Rondon, M. R., August, P. R., Bettermann, A. D., Brady, S. F., Grossman, T. H., Liles, M. R., Loiacono, K. A., Lynch, B. A., MacNeil, I. A. & Minor, C. (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms, *Applied Environmental Microbiology.* **66**, 2541-2547.
24. Bayer, S., Birkemeyer, C. & Ballschmiter, M. (2011) A nitrilase from a metagenomic library acts regioselectively on aliphatic dinitriles, *Applied microbiology and biotechnology.* **89**, 91-98.
25. Wang, K., Li, G., Yu, S. Q., Zhang, C. T. & Liu, Y. H. (2010) A novel metagenome-derived β-galactosidase: gene cloning, overexpression, purification and characterization, *Applied microbiology and biotechnology.* **88**, 155-165.
26. Jiang, C., Li, S.-X., Luo, F.-F., Jin, K., Wang, Q., Hao, Z.-Y., Wu, L.-L., Zhao, G.-C., Ma, G.-F. & Shen, P.-H. (2011) Biochemical characterization of two novel β-glucosidase genes by metagenome expression cloning, *Bioresource technology.* **102**, 3272-3278.
27. Jiang, C., Shen, P., Yan, B. & Wu, B. (2009) Biochemical characterization of a metagenome-derived decarboxylase, *Enzyme and Microbial Technology.* **45**, 58-63.
28. Kotik, M., Štěpánek, V., Grulich, M., Kyslík, P. & Archelas, A. (2010) Access to enantiopure aromatic epoxides and diols using epoxide hydrolases derived from total biofilter DNA, *Journal of Molecular Catalysis B: Enzymatic.* **65**, 41-48.
29. Tiwari, R., Nain, L., Labrou, N. E. & Shukla, P. (2018) Bioprospecting of functional cellulases from metagenome for second generation biofuel production: a review, *Critical Reviews in Microbiology.* **44**, 244-257.
30. Vester, J. K., Glaring, M. A. & Stougaard, P. (2015) Improved cultivation and metagenomics as new tools for bioprospecting in cold environments, *Extremophiles.* **19**, 17-29.
31. Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A. S. & Yandell, M. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes, *Genome research.* **18**, 188-196.
32. Holt, C. & Yandell, M. (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects, *BMC bioinformatics.* **12**, 491.
33. Thibaud-Nissen, F., Souvorov, A., Murphy, T., DiCuccio, M. & Kitts, P. (2013) Eukaryotic genome annotation pipeline in *The NCBI Handbook [Internet] 2nd edition*, Bethesda (MD): National Center for Biotechnology Information (US).
34. Seemann, T. (2014) Prokka: rapid prokaryotic genome annotation, *Bioinformatics.* **30**, 2068-2069.

35. Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M. & Ostell, J. (2016) NCBI prokaryotic genome annotation pipeline, *Nucleic acids research.* **44**, 6614-6624.
36. Chan, K.-L., Rosli, R., Tatarinova, T. V., Hogan, M., Firdaus-Raih, M. & Low, E.-T. L. (2017) Seqping: gene prediction pipeline for plant genomes using self-training gene models and transcriptomic data, *BMC bioinformatics.* **18**, 1.
37. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) Basic local alignment search tool, *Journal of molecular biology.* **215**, 403-410.
38. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic acids research.* **25**, 3389-3402.
39. Eddy, S. R. (2009) A new generation of homology search tools based on probabilistic inference in *Genome Informatics 2009: Genome Informatics Series Vol 23* pp. 205-211, World Scientific.
40. Ponting, C. P., Schultz, J., Copley, R. R., Andrade, M. A. & Bork, P. (2000) Evolution of domain families, *Advances in protein chemistry.* **54**, 185-244.
41. Jacob, F. (2001) Complexity and tinkering, *Annals of the New York Academy of Sciences.* **929**, 71-73.
42. Letunic, I. & Bork, P. (2017) 20 years of the SMART protein domain annotation resource, *Nucleic acids research.* **46**, D493-D496.
43. Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M. & Sangrador-Vegas, A. (2015) The Pfam protein families database: towards a more sustainable future, *Nucleic acids research.* **44**, D279-D285.
44. Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., Brown, S. D., Chang, H.-Y., El-Gebali, S. & Fraser, M. I. (2018) InterPro in 2019: improving coverage, classification and access to protein sequence annotations, *Nucleic acids research.* **47**, D351-D360.
45. Lees, J. G., Lee, D., Studer, R. A., Dawson, N. L., Sillitoe, I., Das, S., Yeats, C., Dessailly, B. H., Rentzsch, R. & Orengo, C. A. (2013) Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis, *Nucleic acids research.* **42,** D240-D245.
46. Doerks, T., Bairoch, A. & Bork, P. (1998) Protein annotation: detective work for function prediction, *Trends in Genetics.* **14**, 248-250.
47. Khersonsky, O., Roodveldt, C. & Tawfik, D. S. (2006) Enzyme promiscuity: evolutionary and mechanistic aspects, *Current opinion in chemical biology.* **10**, 498-508.
48. Coutinho, P. M., Stam, M., Blanc, E. & Henrissat, B. (2003) Why are there so many carbohydrate-active enzyme-related genes in plants?, *Trends in plant science.* **8**, 563-565.
49. Tawfik, D. S. (2010) Messy biology and the origins of evolutionary innovations, *Nature chemical biology.* **6**, 692.
50. Tawfik, O. K. a. D. S. (2010) Enzyme promiscuity: a mechanistic and evolutionary perspective, *Annual review of biochemistry.* **79**, 471-505.
51. Galperin, M. Y. & Koonin, E. V. (2012) Divergence and convergence in enzyme evolution, *Journal of Biological Chemistry.* **287**, 21-28.
52. Pandya, C., Farelli, J. D., Dunaway-Mariano, D. & Allen, K. N. (2014) Enzyme promiscuity: engine of evolutionary innovation, *Journal of Biological Chemistry.* **289**, 30229-30236.

53. Jones, C. E., Brown, A. L. & Baumann, U. (2007) Estimating the annotation error rate of curated GO database sequence annotations, *BMC bioinformatics.* **8**, 170.
54. Gilks, W. R., Audit, B., De Angelis, D., Tsoka, S. & Ouzounis, C. A. (2002) Modeling the percolation of annotation errors in a database of protein sequences, *Bioinformatics.* **18**, 1641-1649.
55. Litthauer, D., Abbai, N. S., Piater, L. A. & van Heerden, E. (2010) Pitfalls using tributyrin agar screening to detect lipolytic activity in metagenomic studies, *African Journal of Biotechnology.* **9**, 4282-4285.
56. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997) A genomic perspective on protein families, *Science.* **278**, 631-637.
57. Lees, J. G., Dawson, N. L., Sillitoe, I. & Orengo, C. A. (2016) Functional innovation from changes in protein domains and their combinations, *Current opinion in structural biology.* **38**, 44-52.
58. Starr, T. N. & Thornton, J. W. (2016) Epistasis in protein evolution, *Protein Science.* **25**, 1204-1218.
59. Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T. & Ben-Tal, N. (2016) ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules, *Nucleic acids research.* **44**, W344-W350.
60. Sydykova, D. K., Jack, B. R., Spielman, S. J. & Wilke, C. O. (2017) Measuring evolutionary rates of proteins in a structural context, *F1000Research.* **6**.
61. Ettema, T. J., de Vos, W. M. & van der Oost, J. (2005) Discovering novel biology by in silico archaeology, *Nature Reviews Microbiology.* **3**, 859-869.
62. Huynen, M. A. & Bork, P. (1998) Measuring genome evolution, *Proceedings of the National Academy of Sciences.* **95**, 5849-5856.
63. Trudeau, D. L. & Tawfik, D. S. (2019) Protein engineers turned evolutionists—the quest for the optimal starting point, *Current opinion in biotechnology.* **60**, 46-52.
64. Balaji, S. & Srinivasan, N. (2007) Comparison of sequence-based and structure-based phylogenetic trees of homologous proteins: Inferences on protein evolution, *Journal of biosciences.* **32**, 83-96.
65. Herman, J. L. (2019) Enhancing Statistical Multiple Sequence Alignment and Tree Inference Using Structural Information in *Computational Methods in Protein Evolution* pp. 183-214, Springer.
66. Lai, J., Jin, J., Kubelka, J. & Liberles, D. A. (2012) A phylogenetic analysis of normal modes evolution in enzymes and its relationship to enzyme function, *Journal of molecular biology.* **422**, 442-459.
67. Khersonsky, O., Lipsh, R., Avizemer, Z., Ashani, Y., Goldsmith, M., Leader, H., Dym, O., Rogotner, S., Trudeau, D. L. & Prilusky, J. (2018) Automated design of efficient and functionally diverse enzyme repertoires, *Molecular cell.* **72**, 178-186.e5.
68. Piel, J., Hui, D., Fusetani, N. & Matsunaga, S. (2004) Targeting modular polyketide synthases with iteratively acting acyltransferases from metagenomes of uncultured bacterial consortia, *Environmental Microbiology.* **6**, 921-927.
69. Elias, M. & Tawfik, D. S. (2012) Divergence and convergence in enzyme evolution: parallel evolution of paraoxonases from quorum-quenching lactonases, *Journal of Biological Chemistry.* **287**, 11-20.

70. Borràs, E., Albalat, R., Duester, G., Parés, X. & Farrés, J. (2014) The Xenopus alcohol dehydrogenase gene family: characterization and comparative analysis incorporating amphibian and reptilian genomes, *BMC genomics.* **15**, 216.
71. Mewis, K., Lenfant, N., Lombard, V. & Henrissat, B. (2016) Dividing the large glycoside hydrolase family 43 into subfamilies: a motivation for detailed enzyme characterization, *Appl Environ Microbiol.* **82**, 1686-1692.
72. Holt, S. M. (2017) Comparative Analysis of Alternansucrase Genes from Leuconostoc Strains, *Transactions of the Illinois State Academy of Science.* **110**, 9-15.
73. van Loo, B., Schober, M., Valkov, E., Heberlein, M., Bornberg-Bauer, E., Faber, K., Hyvönen, M. & Hollfelder, F. (2018) Structural and Mechanistic Analysis of the Choline Sulfatase from Sinorhizobium melliloti: A Class I Sulfatase Specific for an Alkyl Sulfate Ester, *Journal of molecular biology.* **430**, 1004-1023.
74. van Loo, B., Bayer, C. D., Fischer, G., Jonas, S., Valkov, E., Mohamed, M. F., Vorobieva, A., Dutruel, C., Hyvönen, M. & Hollfelder, F. (2018) Balancing specificity and promiscuity in enzyme evolution: multidimensional activity transitions in the alkaline phosphatase superfamily, *Journal of the American Chemical Society.* **141**, 370-387.
75. Laine, R. A. (1994) A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05 x 10 (12) structures for a reducing hexasaccharide: the Isomer Barrier to development of single-method saccharide sequencing or synthesis systems, *Glycobiology.* **4**, 759-767.
76. Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V. & Henrissat, B. (2008) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics, *Nucleic acids research.* **37**, D233-D238.
77. Henrissat, B., Claeyssens, M., Tomme, P., Lemesle, L. & Mornon, J.-P. (1989) Cellulase families revealed by hydrophobic cluster analysi, *Gene.* **81**, 83-95.
78. Couto, F. M., Silva, M. J. & Coutinho, P. (2003). ProFAL: PROtein Functional Annotation through Literature. Paper presented at the JISBD (pp. 747-756).
79. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. (2013) The carbohydrate-active enzymes database (CAZy) in 2013, *Nucleic acids research.* **42**, D490-D495.
80. Stam, M. R., Danchin, E. G., Rancurel, C., Coutinho, P. M. & Henrissat, B. (2006) Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of α-amylase-related proteins, *Protein Engineering, Design and Selection.* **19**, 555-562.
81. St John, F. J., González, J. M. & Pozharski, E. (2010) Consolidation of glycosyl hydrolase family 30: a dual domain 4/7 hydrolase family consisting of two structurally distinct groups, *FEBS letters.* **584**, 4435-4441.
82. Aspeborg, H., Coutinho, P. M., Wang, Y., Brumer, H. & Henrissat, B. (2012) Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5), *BMC evolutionary biology.* **12**, 186.
83. Levasseur, A., Drula, E., Lombard, V., Coutinho, P. M. & Henrissat, B. (2013) Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes, *Biotechnology for biofuels.* **6**, 41.

84. Talamantes, D., Biabini, N., Dang, H., Abdoun, K. & Berlemont, R. (2016) Natural diversity of cellulases, xylanases, and chitinases in bacteria, *Biotechnology for biofuels.* **9**, 133.
85. Berlemont, R. (2017) Distribution and diversity of enzymes for polysaccharide degradation in fungi, *Scientific reports.* **7**, 222.
86. Ravachol, J., Borne, R., Tardif, C., de Philip, P. & Fierobe, H.-P. (2014) Characterization of all family-9 glycoside hydrolases synthesized by the cellulosome-producing bacterium Clostridium cellulolyticum, *Journal of Biological Chemistry.* **289**, 7335-7348.
87. Treseder, K. K. & Lennon, J. T. (2015) Fungal traits that drive ecosystem dynamics on land, *Microbiology and molecular biology reviews.* **79**, 243-262.
88. Llado, S., López-Mondéjar, R. & Baldrian, P. (2017) Forest soil bacteria: diversity, involvement in ecosystem processes, and response to global change, *Microbiol Mol Biol Rev.* **81**, e00063-16.
89. Nguyen, S. T., Freund, H. L., Kasanjian, J. & Berlemont, R. (2018) Function, distribution, and annotation of characterized cellulases, xylanases, and chitinases from CAZy, *Applied microbiology and biotechnology.* **102**, 1629-1637.
90. Helbert, W., Poulet, L., Drouillard, S., Mathieu, S., Loiodice, M., Couturier, M., Lombard, V., Terrapon, N., Turchetto, J. & Vincentelli, R. (2019) Discovery of novel carbohydrate-active enzymes through the rational exploration of the protein sequences space, *Proceedings of the National Academy of Sciences.* **116**, 6063-6068.
91. Devos, D. & Valencia, A. (2000) Practical limits of function prediction, *Proteins: Structure, Function, and Bioinformatics.* **41**, 98-107.
92. Copp, J. N., Akiva, E., Babbitt, P. C. & Tokuriki, N. (2018) Revealing unexplored sequence-function space using sequence similarity networks, *Biochemistry.* **57**, 4651-4662.
93. Atkinson, H. J., Morris, J. H., Ferrin, T. E. & Babbitt, P. C. (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies, *PloS one.* **4**, e4345.
94. Baier, F., Copp, J. & Tokuriki, N. (2016) Evolution of enzyme superfamilies: comprehensive exploration of sequence–function relationships, *Biochemistry.* **55**, 6375-6388.
95. Davies, G. & Henrissat, B. (1995) Structures and mechanisms of glycosyl hydrolases, *Structure.* **3**, 853-859.
96. McCarter, J. D. & Withers, G. S. (1994) Mechanisms of enzymatic glycoside hydrolysis, *Current opinion in structural biology.* **4**, 885-892.
97. Terwisscha van Scheltinga, A. C., Armand, S., Kalk, K. H., Isogai, A., Henrissat, B. & Dijkstra, B. W. (1995) Stereochemistry of chitin hydrolysis by a plant chitinase/lysozyme and x-ray structure of a complex with allosamidin evidence for substrate assisted catalysis, *Biochemistry.* **34**, 15619-15623.
98. Rajan, S. S., Yang, X., Collart, F., Yip, V. L., Withers, S. G., Varrot, A., Thompson, J., Davies, G. J. & Anderson, W. F. (2004) Novel catalytic mechanism of glycoside hydrolysis based on the structure of an NAD+/Mn2+-dependent phospho-α-glucosidase from Bacillus subtilis, *Structure.* **12**, 1619-1629.
99. Liu, Q. P., Sulzenbacher, G., Yuan, H., Bennett, E. P., Pietz, G., Saunders, K., Spence, J., Nudelman, E., Levery, S. B. & White, T. (2007) Bacterial glycosidases for the production of universal red blood cells, *Nature biotechnology.* **25**, 454-464.

100. Rouvinen, J., Bergfors, T., Teeri, T., Knowles, J. & Jones, T. (1990) Three-dimensional structure of cellobiohydrolase II from Trichoderma reesei, *Science.* **249**, 380-386.
101. Divne, C., Stahlberg, J., Reinikainen, T., Ruohonen, L., Pettersson, G., Knowles, J., Teeri, T. T. & Jones, T. A. (1994) The three-dimensional crystal structure of the catalytic core of cellobiohydrolase I from Trichoderma reesei, *Science.* **265**, 524-528.
102. Tharanathan, R. N. & Kittur, F. S. (2003) Chitin—the undisputed biomolecule of great potential l, *Critical reviews in food science and nutrition*. **43**, 61-87.
103. Adrangi, S. & Faramarzi, M. A. (2013) From bacteria to human: a journey into the world of chitinases, *Biotechnology advances.* **31**, 1786-1795.
104. Oyeleye, A. & Normi, Y. M. (2018) Chitinase: diversity, limitations, and trends in engineering for suitable applications, *Bioscience reports.* **38**, BSR2018032300.
105. Vollmer, W., Blanot, D. & De Pedro, M. A. (2008) Peptidoglycan structure and architecture, *FEMS microbiology reviews.* **32**, 149-167.
106. Wohlkönig, A., Huet, J., Looze, Y. & Wintjens, R. (2010) Structural relationships in the lysozyme superfamily: significant evidence for glycoside hydrolase signature motifs, *PloS one.* **5**, e15388.
107. Callewaert, L. & Michiels, C. W. (2010) Lysozymes in the animal kingdom, *Journal of biosciences.* **35**, 127-160.
108. Payne, K. M. & Hatfull, G. F. (2012) Mycobacteriophage endolysins: diverse and modular enzymes with multiple catalytic activities, *PLoS One.* **7**, e34052.
109. Iseli, B., Armand, S., Boller, T., Neuhaus, J.-M. & Henrissat, B. (1996) Plant chitinases use two different hydrolytic mechanisms, *FEBS letters.* **382**, 186-188.
110. Brameld, K. A. & Goddard, W. A. (1998) The role of enzyme distortion in the single displacement mechanism of family 19 chitinases, *Proceedings of the National Academy of Sciences.* **95**, 4276-4281.
111. Ohnuma, T., Umemoto, N., Nagata, T., Shinya, S., Numata, T., Taira, T. & Fukamizo, T. (2014) Crystal structure of a "loopless" GH19 chitinase in complex with chitin tetrasaccharide spanning the catalytic center, *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics.* **1844**, 793-802.
112. Oliveira, H., Vilas Boas, D., Mesnage, S., Kluskens, L. D., Lavigne, R., Sillankorva, S., Secundo, F. & Azeredo, J. (2016) Structural and enzymatic characterization of ABgp46, a novel phage endolysin with broad anti-Gram-negative bacterial activity, *Frontiers in microbiology.* **7**, 208.
113. Lim, J.-A., Shin, H., Kang, D.-H. & Ryu, S. (2012) Characterization of endolysin from a Salmonella Typhimurium-infecting bacteriophage SPN1S, *Research in microbiology.* **163**, 233-241.
114. Prakash, N. U., Jayanthi, M., Sabarinathan, R., Kangueane, P., Mathew, L. & Sekar, K. (2010) Evolution, homology conservation, and identification of unique sequence signatures in GH19 family chitinases, *Journal of molecular evolution.* **70**, 466-478.
115. Loessner, M. J. (2005) Bacteriophage endolysins—current state of research and applications, *Current opinion in microbiology.* **8**, 480-487.
116. Boller, T. (1987) Hydrolytic enzymes in plant disease resistance, *Plant-microbe interactions, molecular and genetic perspectives.* **3**, 385-411.

117. Legrand, M., Kauffmann, S., Geoffroy, P. & Fritig, B. (1987) Biological function of pathogenesis-related proteins: four tobacco pathogenesis-related proteins are chitinases, *Proceedings of the National Academy of Sciences.* **84**, 6750-6754.
118. Shinshi, H., Neuhaus, J.-M., Ryals, J. & Meins, F. (1990) Structure of a tobacco endochitinase gene: evidence that different chitinase genes can arise by transposition of sequences encoding a cysteine-rich domain, *Plant molecular biology.* **14**, 357-368.
119. Collinge, D. B., Kragh, K. M., Mikkelsen, J. D., Nielsen, K. K., Rasmussen, U. & Vad1, K. (1993) Plant chitinases, *The plant Journal.* **3**, 31-40.
120. Tanaka, J., Fukamizo, T. & Ohnuma, T. (2017) Enzymatic properties of a GH19 chitinase isolated from rice lacking a major loop structure involved in chitin binding, *Glycobiology.* **27**, 477-485.
121. Ubhayasekera, W., Rawat, R., Ho, S. W. T., Wiweger, M., Von Arnold, S., Chye, M.-L. & Mowbray, S. L. (2009) The first crystal structures of a family 19 class IV chitinase: the enzyme from Norway spruce, *Plant molecular biology.* **71**, 277-289.
122. Taira, T., Mahoe, Y., Kawamoto, N., Onaga, S., Iwasaki, H., Ohnuma, T. & Fukamizo, T. (2011) Cloning and characterization of a small family 19 chitinase from moss (Bryum coronatum), *Glycobiology.* **21**, 644-654.
123. Ohnuma, T., Sørlie, M., Fukuda, T., Kawamoto, N., Taira, T. & Fukamizo, T. (2011) Chitin oligosaccharide binding to a family GH19 chitinase from the moss Bryum coronatum, *The FEBS journal.* **278**, 3991-4001.
124. Watanabe, T., Kanai, R., Kawase, T., Tanabe, T., Mitsutomi, M., Sakuda, S. & Miyashita, K. (1999) Family 19 chitinases of Streptomyces species: characterization and distribution, *Microbiology.* **145**, 3353-3363.
125. Hoell, I. A., Dalhus, B., Heggset, E. B., Aspmo, S. I. & Eijsink, V. G. (2006) Crystal structure and enzymatic properties of a bacterial family 19 chitinase reveal differences from plant enzymes, *The FEBS journal.* **273**, 4889-4900.
126. Kawase, T., Saito, A., Sato, T., Kanai, R., Fujii, T., Nikaidou, N., Miyashita, K. & Watanabe, T. (2004) Distribution and phylogenetic analysis of family 19 chitinases in Actinobacteria, *Applied and environmental microbiology.* **70**, 1135-1144.
127. Shimosaka, M., Fukumori, Y., Narita, T., Zhang, X.-Y., Kodaira, R., Nogawa, M. & Okazaki, M. (2001) The bacterium Burkholderia gladioli strain CHB101 produces two different kinds of chitinases belonging to families 18 and 19 of the glycosyl hydrolases, *Journal of bioscience and bioengineering.* **91**, 103-105.
128. Ueda, M., Kojima, M., Yoshikawa, T., Mitsuda, N., Araki, K., Kawaguchi, T., Miyatake, K., Arai, M. & Fukamizo, T. (2003) A novel type of family 19 chitinase from Aeromonas sp. No. 10S-24: Cloning, sequence, expression, and the enzymatic properties, *European journal of biochemistry.* **270**, 2513-2520.
129. Honda, Y., Taniguchi, H. & Kitaoka, M. (2008) A reducing-end-acting chitinase from Vibrio proteolyticus belonging to glycoside hydrolase family 19, *Applied microbiology and biotechnology.* **78**, 627-634.
130. García-Fraga, B., da Silva, A. F., López-Seijas, J. & Sieiro, C. (2015) A novel family 19 chitinase from the marine-derived Pseudoalteromonas tunicata CCUG 44952T: Heterologous expression, characterization and antifungal activity, *Biochemical engineering journal.* **93**, 84-93.

131. Martínez-Caballero, S., Cano-Sánchez, P., Mares-Mejía, I., Díaz-Sánchez, A. G., Macías-Rubalcava, M. L., Hermoso, J. A. & Rodríguez-Romero, A. (2014) Comparative study of two GH19 chitinase-like proteins from Hevea brasiliensis, one exhibiting a novel carbohydrate-binding domain, *The FEBS journal.* **281**, 4535-4554.
132. Saul, F. A., Rovira, P., Boulot, G., Van Damme, E. J., Peumans, W. J., Truffa-Bachi, P. & Bentley, G. A. (2000) Crystal structure of Urtica dioica agglutinin, a superantigen presented by MHC molecules of class I and class II, *Structure.* **8**, 593-603.
133. Lai, M.-J., Lin, N.-T., Hu, A., Soo, P.-C., Chen, L.-K., Chen, L.-H. & Chang, K.-C. (2011) Antibacterial activity of Acinetobacter baumannii phage ϕAB2 endolysin (LysAB2) against both gram-positive and gram-negative bacteria, *Applied microbiology and biotechnology.* **90**, 529-539.
134. Peng, S.-Y., You, R.-I., Lai, M.-J., Lin, N.-T., Chen, L.-K. & Chang, K.-C. (2017) Highly potent antimicrobial modified peptides derived from the Acinetobacter baumannii phage endolysin LysAB2, *Scientific reports.* **7**, 11477.
135. Pohane, A. A., Joshi, H. & Jain, V. (2014) Molecular dissection of phage endolysin: an interdomain interaction confers host specificity in Lysin A of Mycobacterium phage D29, *Journal of biological chemistry.* **289**, 12085-12095.
136. Walmagh, M., Briers, Y., Dos Santos, S. B., Azeredo, J. & Lavigne, R. (2012) Characterization of modular bacteriophage endolysins from Myoviridae phages OBP, 201ϕ2-1 and PVP-SE1, *PLoS One.* **7**, e36991.
137. Hosoda, N., Kurokawa, Y., Sako, Y., Nagasaki, K., Yoshida, T. & Hiroishi, S. (2011) The functional effect of Gly209 and Ile213 substitutions on lysozyme activity of family 19 chitinase encoded by cyanophage Ma-LMM01, *Fisheries Science.* **77**, 665-670.
138. Yamada, T., Satoh, S., Ishikawa, H., Fujiwara, A., Kawasaki, T., Fujie, M. & Ogata, H. (2010) A jumbo phage infecting the phytopathogen Ralstonia solanacearum defines a new lineage of the Myoviridae family, *Virology.* **398**, 135-147.
139. Ohnuma, T., Umemoto, N., Kondo, K., Numata, T. & Fukamizo, T. (2013) Complete subsite mapping of a "loopful" GH19 chitinase from rye seeds based on its crystal structure, *FEBS letters.* **587**, 2691-2697.
140. Davies, G. J., Wilson, K. S. & Henrissat, B. (1997) Nomenclature for sugar-binding subsites in glycosyl hydrolases, *Biochemical Journal.* **321**, 557-559.
141. Park, Y., Lim, J. A., Kong, M., Ryu, S. & Rhee, S. (2014) Structure of bacteriophage SPN 1 S endolysin reveals an unusual two-module fold for the peptidoglycan lytic and binding activity, *Molecular microbiology.* **92**, 316-325.
142. Boller, T. (1988) Ethylene and the regulation of antifungal hydrolases in plants, *Oxford Surveys of Plant Molecular and Cell Biology (United Kingdom)*.
143. Zhu, Q. & Lamb, C. J. (1991) Isolation and characterization of a rice gene encoding a basic chitinase, *Molecular and general genetics MGG.* **226**, 289-296.
144. Leah, R., Tommerup, H., Svendsen, I. & Mundy, J. (1991) Biochemical and molecular characterization of three barley seed proteins with antifungal properties, *Journal of biological Chemistry.* **266**, 1564-1573.
145. Velazhahan, R., Datta, S. & Muthukrishnan, S. (1999) Plant Chitinases (PR-3, PR-4, PR-8, PR-11) in *Pathogenesis-related proteins in plants*, CRC Press.

146. Ebrahim, S., Usha, K. & Singh, B. (2011) Pathogenesis related (PR) proteins in *Plant defense mechanism in Science against microbial pathogens: communicating current research and technological advances* pp. 1043-1054, Formatex Research Center.
147. Rawat, S., Ali, S., Mittra, B. & Grover, A. (2017) Expression analysis of chitinase upon challenge inoculation to Alternaria wounding and defense inducers in *Brassica juncea*, *Biotechnology reports.* **13**, 72-79.
148. Fernandes, J. C., Tavaria, F. K., Soares, J. C., Ramos, Ó. S., Monteiro, M. J., Pintado, M. E. & Malcata, F. X. (2008) Antimicrobial effects of chitosans and chitooligosaccharides, upon Staphylococcus aureus and Escherichia coli, in food model systems, *Food microbiology.* **25**, 922-928.
149. Aam, B. B., Heggset, E. B., Norberg, A. L., Sørlie, M., Vårum, K. M. & Eijsink, V. G. (2010) Production of chitooligosaccharides and their potential applications in medicine, *Marine drugs.* **8**, 1482-1517.
150. Liaqat, F. & Eltem, R. (2018) Chitooligosaccharides and their biological activities: A comprehensive review, *Carbohydrate polymers.* **184**, 243-259.
151. Zou, P., Yuan, S., Yang, X., Zhai, X. & Wang, J. (2018) Chitosan oligosaccharides with degree of polymerization 2–6 induces apoptosis in human colon carcinoma HCT116 cells, *Chemico-biological interactions.* **279**, 129-135.
152. Halder, S. K., Pal, S. & Mondal, K. C. (2019) Biosynthesis of Fungal Chitinolytic Enzymes and Their Potent Biotechnological Appliances in *Recent Advancement in White Biotechnology Through Fungi* pp. 281-298, Springer.
153. Kumari, S., Rath, P., Kumar, A. S. H. & Tiwari, T. (2015) Extraction and characterization of chitin and chitosan from fishery waste by chemical method, *Environmental technology & innovation.* **3**, 77-85.
154. Le, B. & Yang, S. H. (2019) Microbial chitinases: properties, current state and biotechnological applications, *World Journal of microbiology and biotechnology.* **35**, 144.
155. Ohnuma, T., Tanaka, T., Urasaki, A., Dozen, S. & Fukamizo, T. (2018) A novel method for chemo-enzymatic synthesis of chitin oligosaccharide catalyzed by the mutant of inverting family GH19 chitinase using 4, 6-dimethoxy-1, 3, 5-triazin-2-yl α-chitobioside as a glycosyl donor, *The journal of biochemistry.* **165**, 497-503.
156. Schmelcher, M., Donovan, D. M. & Loessner, M. J. (2012) Bacteriophage endolysins as novel antimicrobials, *Future microbiology.* **7**, 1147-1171.
157. Briers, Y., Walmagh, M., Van Puyenbroeck, V., Cornelissen, A., Cenens, W., Aertsen, A., Oliveira, H., Azeredo, J., Verween, G. & Pirnay, J.-P. (2014) Engineered endolysin-based "Artilysins" to combat multidrug-resistant gram-negative pathogens, *MBio.* **5**, e01379-14.
158. Gerstmans, H., Rodriguez-Rubio, L., Lavigne, R. & Briers, Y. (2016) From endolysins to Artilysin® s: novel enzyme-based approaches to kill drug-resistant bacteria, *Biochemical Society Transactions.* **44**, 123-128.
159. Latka, A., Maciejewska, B., Majkowska-Skrobek, G., Briers, Y. & Drulis-Kawa, Z. (2017) Bacteriophage-encoded virion-associated enzymes to overcome the carbohydrate barriers during the infection process, *Applied microbiology and biotechnology.* **101**, 3103-3119.
160. Rathore, A. S. & Gupta, R. D. (2015) Chitinases from bacteria to human: properties, applications, and future perspectives, *Enzyme research.* **2015**, 791907.

161. D'Amico, S., Collins, T., Marx, J. C., Feller, G. & Gerday, C. (2006) Psychrophilic microorganisms: challenges for life, *EMBO reports.* **7**, 385-389.
162. De Maayer, P., Anderson, D., Cary, C. & Cowan, D. A. (2014) Some like it cold: understanding the survival strategies of psychrophiles, *EMBO reports.* **15**, 508-517.
163. Morita, R. Y. (1975) Psychrophilic bacteria, *Bacteriological reviews.* **39**, 144.
164. Mazur, P. (1984) Freezing of living cells: mechanisms and implications, *American journal of physiology-cell physiology.* **247**, C125-C142.
165. Chintalapati, S., Kiran, M. & Shivaji, S. (2004) Role of membrane lipid fatty acids in cold adaptation, *Cellular and molecular biology (Noisy-le-Grand, France).* **50**, 631-642.
166. Pischedda, A., Ramasamy, K. P., Mangiagalli, M., Chiappori, F., Milanesi, L., Miceli, C., Pucciarelli, S. & Lotti, M. (2018) Antarctic marine ciliates under stress: superoxide dismutases from the psychrophilic Euplotes focardii are cold-active yet heat tolerant enzymes, *Scientific reports.* **8**, 14721.
167. Laidler, K. J. (1984) The development of the Arrhenius equation, *Journal of chemical education.* **61**, 494.
168. Lonhienne, T., Zoidakis, J., Vorgias, C. E., Feller, G., Gerday, C. & Bouriotis, V. (2001) Modular structure, local flexibility and cold-activity of a novel chitobiase from a psychrophilic Antarctic bacterium, *Journal of molecular biology.* **310**, 291-297.
169. Matsuura, A., Yao, M., Aizawa, T., Koganesawa, N., Masaki, K., Miyazawa, M., Demura, M., Tanaka, I., Kawano, K. & Nitta, K. (2002) Structural analysis of an insect lysozyme exhibiting catalytic efficiency at low temperatures, *Biochemistry.* **41**, 12086-12092.
170. D'Amico, S., Gerday, C. & Feller, G. (2002) Dual effects of an extra disulfide bond on the activity and stability of a cold-adapted α-amylase, *Journal of biological chemistry.* **277**, 46110-46115.
171. D'Amico, S., Gerday, C. & Feller, G. (2003) Temperature adaptation of proteins: engineering mesophilic-like activity and stability in a cold-adapted α-amylase, *Journal of molecular biology.* **332**, 981-988.
172. D'Amico, S., Marx, J.-C., Gerday, C. & Feller, G. (2003) Activity-stability relationships in extremophilic enzymes, *Journal of biological chemistry.* **278**, 7891-7896.
173. Mavromatis, K., Lorito, M., Woo, S. L. & Bouriotis, V. (2003) Mode of action and antifungal properties of two cold-adapted chitinases, *Extremophiles.* **7**, 385-390.
174. Liang, Z.-X., Tsigos, I., Lee, T., Bouriotis, V., Resing, K. A., Ahn, N. G. & Klinman, J. P. (2004) Evidence for increased local flexibility in psychrophilic alcohol dehydrogenase relative to its thermophilic homologue, *Biochemistry.* **43**, 14676-14683.
175. Fedøy, A.-E., Yang, N., Martinez, A., Leiros, H.-K. S. & Steen, I. H. (2007) Structural and functional properties of isocitrate dehydrogenase from the psychrophilic bacterium Desulfotalea psychrophila reveal a cold-active enzyme with an unusual high thermal stability, *Journal of molecular biology.* **372**, 130-149.
176. Riise, E. K., Lorentzen, M. S., Helland, R., Smalås, A. O., Leiros, H.-K. & Willassen, N. P. (2007) The first structure of a cold-active catalase from Vibrio salmonicida at 1.96 Å reveals structural aspects of cold adaptation, *Acta crystallographica section D: biological crystallography.* **63**, 135-148.

177. Lian, K., Leiros, H.-K. S. & Moe, E. (2015) MutT from the fish pathogen Aliivibrio salmonicida is a cold-active nucleotide-pool sanitization enzyme with unexpectedly high thermostability, *FEBS open bio.* **5**, 107-116.
178. Eyring, H. (1935) The activated complex and the absolute rate of chemical reactions, *Chemical reviews.* **17**, 65-77.
179. Lonhienne, T., Gerday, C. & Feller, G. (2000) Psychrophilic enzymes: revisiting the thermodynamic parameters of activation may explain local flexibility, *Biochimica et biophysica acta (BBA)-protein structure and molecular enzymology.* **1543**, 1-10.
180. Siddiqui, K. S. & Cavicchioli, R. (2006) Cold-adapted enzymes, *Annual review of biochemistry.* **75**, 403-433.
181. Feller, G. & Gerday, C. (1997) Psychrophilic enzymes: molecular basis of cold adaptation, *Cellular and molecular life sciences CMLS.* **53**, 830-841.
182. Åqvist, J., Isaksen, G. V. & Brandsdal, B. O. (2017) Computation of enzyme cold adaptation, *Nature reviews chemistry.* **1**, 0051.
183. Mangiagalli, M., Brocca, S., Orlando, M. & Lotti, M. (2019) The "cold revolution". Present and future applications of cold-active enzymes and ice-binding proteins, *New biotechnology*. **55**, 5-11.
184. Feller, G. (2013) Psychrophilic enzymes: from folding to function and biotechnology, *Scientifica.* **2013**, 512840.
185. Santiago, M., Ramírez-Sarmiento, C. A., Zamora, R. A. & Parra, L. P. (2016) Discovery, molecular mechanisms, and industrial applications of cold-active enzymes, *Frontiers in microbiology.* **7**, 1408.
186. Cavicchioli, R., Charlton, T., Ertan, H., Omar, S. M., Siddiqui, K. & Williams, T. (2011) Biotechnological uses of enzymes from psychrophiles, *Microbial biotechnology.* **4**, 449-460.
187. Margesin, R., Feller, G., Gerday, C. & Russell, N. J. (2003) Cold-adapted microorganisms: adaptation strategies and biotechnological potential, *Encyclopedia of environmental microbiology*, Wiley Online Library.
188. Cavicchioli, R., Siddiqui, K. S., Andrews, D. & Sowers, K. R. (2002) Low-temperature extremophiles and their applications, *Current opinion in biotechnology.* **13**, 253-261.
189. Valbonesi, A. & Luporini, P. (1993) Biology of Euplotes focardii, an Antarctic ciliate, *Polar biology.* **13**, 489-493.
190. Pucciarelli, S., La Terza, A., Ballarini, P., Barchetta, S., Yu, T., Marziale, F., Passini, V., Methé, B., Detrich III, H. W. & Miceli, C. (2009) Molecular cold-adaptation of protein function and gene regulation: the case for comparative genomic analyses in marine ciliated protozoa, *Marine genomics.* **2**, 57-66.
191. Pucciarelli, S., Devaraj, R. R., Mancini, A., Ballarini, P., Castelli, M., Schrallhammer, M., Petroni, G. & Miceli, C. (2015) Microbial consortium associated with the Antarctic marine ciliate Euplotes focardii: an investigation from genomic sequences, *Microbial ecology.* **70**, 484-497.
192. Ramasamy, K. P., Telatin, A., Mozzicafreddo, M., Miceli, C. & Pucciarelli, S. (2019) Draft Genome Sequence of a New Pseudomonas sp. Strain, ef1, Associated with the Psychrophilic Antarctic Ciliate Euplotes focardii, *Microbiology Resource Announcements.* **8**, e00867-19.

193. Pohane, A. A., Joshi, H. & Jain, V. (2014) Molecular Dissection of Phage Endolysin an interdomain interaction confers host specificity in lysin a of mycobacterium phage D29, *Journal of biological chemistry.* **289**, 12085-12095.
194. Alquati, C., De Gioia, L., Santarossa, G., Alberghina, L., Fantucci, P. & Lotti, M. (2002) The cold-active lipase of Pseudomonas fragi: Heterologous expression, biochemical characterization and molecular modeling, *European Journal of Biochemistry.* **269**, 3321-3328.
195. Rojas-Contreras, J. A., de la Rosa, A. P. B. & De León-Rodríguez, A. (2015) Expression and characterization of a recombinant psychrophilic Cu/Zn superoxide dismutase from Deschampsia antarctica E. Desv.[Poaceae], *Applied biochemistry and biotechnology.* **175**, 3287-3296.
196. Stern, A. & Sorek, R. (2011) The phage-host arms race: shaping the evolution of microbes, *Bioessays.* **33**, 43-51.
197. Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H. & Xu, J. (2012) Template-based protein structure modeling using the RaptorX web server, *Nature protocols.* **7**, 1511-1522.
198. Lacombe-Harvey, M.-È., Brzezinski, R. & Beaulieu, C. (2018) Chitinolytic functions in actinobacteria: ecology, enzymes, and evolution, *Applied microbiology and biotechnology.* **102**, 7219-7230.
199. Leoni, C., Volpicella, M., Dileo, M. C., Gattulli, B. A. & Ceci, L. R. (2019) Chitinases as food allergens, *Molecules.* **24**, 2087.
200. Júnior, J. E. M., Grangeiro, T. B. & Nogueira, N. A. P. (2018) Chitinases as Antibacterial Proteins: A Systematic Review, *Journal of young pharmacists.* **10**, 144-148.
201. Ohnuma, T., Numata, T., Osawa, T., Inanaga, H., Okazaki, Y., Shinya, S., Kondo, K., Fukuda, T. & Fukamizo, T. (2012) Crystal structure and chitin oligosaccharide-binding mode of a 'loopful'family GH19 chitinase from rye, Secale cereale, seeds, *The FEBS journal.* **279**, 3639-3651.
202. Brunner, F., Stintzi, A., Fritig, B. & Legrand, M. (1998) Substrate specificities of tobacco chitinases, *The plant journal.* **14**, 225-234.
203. Taira, T., Ohnuma, T., Yamagami, T., Aso, Y., Ishiguro, M. & Ishihara, M. (2002) Antifungal activity of rye (Secale cereale) seed chitinases: the different binding manner of class I and class II chitinases to the fungal cell walls, *Bioscience, biotechnology, and biochemistry.* **66**, 970-977.
204. Nakamura, S., Iwai, T., Honkura, R., UGAKI, M., OHSHIMA, M. & OHASHI, Y. (1997) Four chitinase cDNAs from Chenopodium amaranticolor, *Plant biotechnology.* **14**, 85-86.
205. Suarez, V., Staehelin, C., Arango, R., Holtorf, H., Hofsteenge, J. & Meins, F. (2001) Substrate specificity and antifungal activity of recombinant tobacco class I chitinases, *Plant molecular biology.* **45**, 609-618.
206. Gerlt, J. A. (2017) Genomic enzymology: web tools for leveraging protein family sequence–function space and genome context to discover novel functions, *Biochemistry.* **56**, 4293-4308.
207. Takenaka, S., Ohnuma, T. & Fukamizo, T. (2017) Insertion of a Loop Structure into the "Loopless" GH19 Chitinase from Bryum coronatum, *Journal of applied glycoscience.* **64**, 39-42.

208. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome research.* **13**, 2498-2504.
209. Hossain, M. A., Noh, H.-N., Kim, K.-I., Koh, E.-J., Wi, S.-G., Bae, H.-J., Lee, H. & Hong, S.-W. (2010) Mutation of the chitinase-like protein-encoding AtCTL2 gene enhances lignin accumulation in dark-grown Arabidopsis seedlings, *Journal of plant physiology.* **167**, 650-658.
210. Kwon, Y., Kim, S. H., Jung, M. S., Kim, M. S., Oh, J. E., Ju, H. W., Kim, K. i., Vierling, E., Lee, H. & Hong, S. W. (2007) Arabidopsis hot2 encodes an endochitinase-like protein that is essential for tolerance to heat, salt and drought stresses, *The plant Journal.* **49**, 184-193.
211. Wasano, N., Konno, K., Nakamura, M., Hirayama, C., Hattori, M. & Tateishi, K. (2009) A unique latex protein, MLX56, defends mulberry trees from insects, *Phytochemistry.* **70**, 880-888.
212. Martínez-Caballero, S., Cano-Sánchez, P., Mares-Mejía, I., Díaz-Sánchez, A. G., Macías-Rubalcava, M. L., Hermoso, J. A. & Rodríguez-Romero, A. (2014) Comparative study of two GH 19 chitinase-like proteins from Hevea brasiliensis, one exhibiting a novel carbohydrate-binding domain, *The FEBS journal.* **281**, 4535-4554.
213. Huet, J., Rucktooa, P., Clantin, B., Azarkan, M., Looze, Y., Villeret, V. & Wintjens, R. (2008) X-ray structure of papaya chitinase reveals the substrate binding mode of glycosyl hydrolase family 19 chitinases, *Biochemistry.* **47**, 8283-8291.
214. Landim, P. G. C., Correia, T. O., Silva, F. D., Nepomuceno, D. R., Costa, H. P., Pereira, H. M., Lobo, M. D., Moreno, F. B., Brandão-Neto, J. & Medeiros, S. C. (2017) Production in Pichia pastoris, antifungal activity and crystal structure of a class I chitinase from cowpea (Vigna unguiculata): Insights into sugar binding mode and hydrolytic action, *Biochimie.* **135**, 89-103.
215. Honda, Y. & Fukamizo, T. (1998) Substrate binding subsites of chitinase from barley seeds and lysozyme from goose egg white, *Biochimica et biophysica acta (BBA)-protein structure and molecular enzymology.* **1388**, 53-65.
216. Sasaki, C., Itoh, Y., Takehara, H., Kuhara, S. & Fukamizo, T. (2003) Family 19 chitinase from rice (Oryza sativa L.): substrate-binding subsites demonstrated by kinetic and molecular modeling studies, *Plant molecular biology.* **52**, 43-52.
217. Chaudet, M. M., Naumann, T. A., Price, N. P. & Rose, D. R. (2014) Crystallographic structure of ChitA, a glycoside hydrolase family 19, plant class IV chitinase from Zea mays, *Protein science.* **23**, 586-593.
218. Fukamizo, T., Miyake, R., Tamura, A., Ohnuma, T., Skriver, K., Pursiainen, N. V. & Juffer, A. H. (2009) A flexible loop controlling the enzymatic activity and specificity in a glycosyl hydrolase family 19 endochitinase from barley seeds (Hordeum vulgare L.), *Biochimica et biophysica acta (BBA)-proteins and proteomics.* **1794**, 1159-1167.
219. Letzel, T., Sahmel-Schneider, E., Skriver, K., Ohnuma, T. & Fukamizo, T. (2011) Chitinase-catalyzed hydrolysis of 4-nitrophenyl penta-N-acetyl-β-chitopentaoside as determined by real-time ESIMS: The 4-nitrophenyl moiety of the substrate interacts with the enzyme binding site, *Carbohydrate research.* **346**, 863-866.
220. Mizuno, R., Fukamizo, T., Sugiyama, S., Nishizawa, Y., Kezuka, Y., Nonaka, T., Suzuki, K. & Watanabe, T. (2008) Role of the loop structure of the catalytic domain in rice class I chitinase, *Journal of biochemistry.* **143**, 487-495.

221. Truong, N.-H., Park, S.-M., Nishizawa, Y., Watanabe, T., Sasaki, T. & Itoh, Y. (2003) Structure, heterologous expression, and properties of rice (Oryza sativa L.) family 19 chitinases, *Bioscience, biotechnology, and biochemistry.* **67**, 1063-1070.
222. Tamura, M., Miyazaki, T., Tanaka, Y., Yoshida, M., Nishikawa, A. & Tonozuka, T. (2012) Comparison of the structural changes in two cellobiohydrolases, CcCel6A and CcCel6C, from Coprinopsis cinerea–a tweezer-like motion in the structure of CcCel6C, *The FEBS journal.* **279**, 1871-1882.
223. Han, B., Zhou, K., Li, Z., Sun, B., Ni, Q., Meng, X., Pan, G., Li, C., Long, M. & Li, T. (2016) Characterization of the first fungal Glycosyl Hydrolase family 19 chitinase (NbchiA) from Nosema bombycis (Nb), *Journal of eukaryotic microbiology.* **63**, 37-45.
224. Schlöffel, M. A., Käsbauer, C. & Gust, A. A. (2019) Interplay of plant glycan hydrolases and LysM proteins in plant–bacteria interactions, *International journal of medical microbiology*. **309**, 252-257.
225. Fan, J., Wang, H., Feng, D., Liu, B., Liu, H. & Wang, J. (2007) Molecular characterization of plantain class I chitinase gene and its expression in response to infection by Gloeosporium musarum Cke and Massee and other abiotic stimuli, *Journal of biochemistry.* **142**, 561-570.
226. López, R. C. & Gómez-Gómez, L. (2009) Isolation of a new fungi and wound-induced chitinase class in corms of Crocus sativus, *Plant physiology and biochemistry.* **47**, 426-434.
227. Buchholz, P. C., Vogel, C., Reusch, W., Pohl, M., Rother, D., Spieß, A. C. & Pleiss, J. (2016) BioCatNet: a database system for the integration of enzyme sequences and biocatalytic experiments, *ChemBioChem.* **17**, 2093-2098.
228. Feller, G. & Gerday, C. (2003) Psychrophilic enzymes: hot topics in cold adaptation, *Nature reviews microbiology.* **1**, 200-208.
229. Siddiqui, K. S., Feller, G., D'Amico, S., Gerday, C., Giaquinto, L. & Cavicchioli, R. (2005) The active site is the least stable structure in the unfolding pathway of a multidomain cold-adapted α-amylase, *Journal of bacteriology.* **187**, 6197-6205.
230. Collins, T. & Margesin, R. (2019) Psychrophilic lifestyles: mechanisms of adaptation and biotechnological tools, *Applied microbiology and biotechnology*. **103**, 1-15.
231. Mangiagalli, M., Bar-Dolev, M., Tedesco, P., Natalello, A., Kaleda, A., Brocca, S., Pascale, D., Pucciarelli, S., Miceli, C. & Braslavsky, I. (2017) Cryo-protective effect of an ice-binding protein derived from Antarctic bacteria, *The FEBS journal.* **284**, 163-177.
232. Oliveira, H., Melo, L. D., Santos, S. B., Nóbrega, F. L., Ferreira, E. C., Cerca, N., Azeredo, J. & Kluskens, L. D. (2013) Molecular aspects and comparative genomics of bacteriophage endolysins, *Journal of virology.* **87**, 4558-4570.
233. Vidová, B., Šramková, Z., Tišáková, L., Oravkinová, M. & Godány, A. (2014) Bioinformatics analysis of bacteriophage and prophage endolysin domains, *Biologia.* **69**, 541-556.
234. Studier, F. W. (2005) Protein production by auto-induction in high-density shaking cultures, *Protein expression and purification.* **41**, 207-234.
235. Brocca, S., Ferrari, C., Barbiroli, A., Pesce, A., Lotti, M. & Nardini, M. (2016) A bacterial acyl aminoacyl peptidase couples flexibility and stability as a result of cold adaptation, *The FEBS journal.* **283**, 4310-4324.
236. Briers, Y., Lavigne, R., Volckaert, G. & Hertveldt, K. (2007) A standardized approach for accurate quantification of murein hydrolase activity in high-throughput assays, *Journal of biochemical and biophysical methods.* **70**, 531-533.

237. Shen, C.-R., Chen, Y.-S., Yang, C.-J., Chen, J.-K. & Liu, C.-L. (2010) Colloid chitin azure is a dispersible, low-cost substrate for chitinase measurements in a sensitive, fast, reproducible assay, *Journal of biomolecular screening.* **15**, 213-217.
238. Armenteros, J. J. A., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., von Heijne, G. & Nielsen, H. (2019) SignalP 5.0 improves signal peptide predictions using deep neural networks, *Nature biotechnology*. **37**, 420-423.
239. Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y. & Wishart, D. S. (2016) PHASTER: a better, faster version of the PHAST phage search tool, *Nucleic acids research.* **44**, W16-W21.
240. Katoh, K. & Standley, D. M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability, *Molecular biology and evolution.* **30**, 772-780.
241. Gascuel, O. (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data, *Molecular biology and evolution.* **14**, 685-695.
242. Suchard, M. A. & Redelings, B. D. (2006) BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny, *Bioinformatics.* **22**, 2047-2048.
243. Le, S. Q. & Gascuel, O. (2008) An improved general amino acid replacement matrix, *Molecular biology and evolution.* **25**, 1307-1320.
244. Redelings, B. D. & Suchard, M. A. (2007) Incorporating indel information into phylogeny estimation for rapidly emerging pathogens, *BMC evolutionary biology.* **7**, 40.
245. Bansal, M. S., Kellis, M., Kordi, M. & Kundu, S. (2018) RANGER-DTL 2.0: Rigorous Reconstruction of Gene-Family Evolution by Duplication, Transfer, and Loss, *Bioinformatics*. **34**, 3214-3216.
246. Gouy, M., Guindon, S. & Gascuel, O. (2009) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building, *Molecular biology and evolution.* **27**, 221-224.
247. Mayrose, I., Graur, D., Ben-Tal, N. & Pupko, T. (2004) Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior, *Molecular biology and evolution.* **21**, 1781-1791.
248. Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T. & Ben-Tal, N. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures, *Nucleic acids research.* **33**, W299-W302.
249. MacKerell Jr, A. D., Bashford, D., Bellott, M., Dunbrack Jr, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H. & Ha, S. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins, *The journal of physical chemistry B.* **102**, 3586-3616.
250. Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B. & Lindahl, E. (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers, *SoftwareX.* **1**, 19-25.
251. Kwon, S. W., Kim, J. S., Park, I. C., Yoon, S. H., Park, D. H., Lim, C. K. & Go, S. J. (2003) Pseudomonas koreensis sp. nov., Pseudomonas umsongensis sp. nov. and Pseudomonas jinjuensis sp. nov., novel species from farm soils in Korea, *International journal of systematic and evolutionary microbiology.* **53**, 21-27.
252. Kelly, S. M., Jess, T. J. & Price, N. C. (2005) How to study proteins by circular dichroism, *Biochimica et biophysica acta (BBA)-proteins and proteomics.* **1751**, 119-139.

253. Yang, G., De Santi, C., de Pascale, D., Pucciarelli, S., Pucciarelli, S. & Miceli, C. (2013) Characterization of the first eukaryotic cold-adapted patatin-like phospholipase from the psychrophilic Euplotes focardii: identification of putative determinants of thermal-adaptation by comparison with the homologous protein from the mesophilic Euplotes crassus, *Biochimie.* **95**, 1795-1806.
254. Yang, G., Yao, H., Mozzicafreddo, M., Ballarini, P., Pucciarelli, S. & Miceli, C. (2017) Rational engineering of a cold-adapted α-amylase from the Antarctic ciliate Euplotes focardii for simultaneous improvement of thermostability and catalytic activity, *Applied environmental microbiology.* **83**, e00449-17.
255. Veiga-Crespo, P., Ageitos, J. M., Poza, M. & Villa, T. G. (2007) Enzybiotics: a look to the future, recalling the past, *Journal of pharmaceutical sciences.* **96**, 1917-1924.
256. Banerjee, S. K., Holler, E., Hess, G. P. & Rupley, J. A. (1975) Reaction of N-acetylglucosamine oligosaccharides with lysozyme. Temperature, pH, and solvent deuterium isotope effects; equilbrium, steady state, and pre-steady state measurements*, *Journal of biological chemistry.* **250**, 4355-4367.
257. Boller, T., Gehri, A., Mauch, F. & Vögeli, U. (1983) Chitinase in bean leaves: induction by ethylene, purification, properties, and possible function, *Planta.* **157**, 22-31.
258. Bokma, E., van Koningsveld, G. A., Jeronimus-Stratingh, M. & Beintema, J. J. (1997) Hevamine, a chitinase from the rubber tree Hevea brasiliensis, cleaves peptidoglycan between the C-1 of N-acetylglucosamine and C-4 of N-acetylmuramic acid and therefore is not a lysozyme, *FEBS letters.* **411**, 161-163.
259. Wang, S.-L. & Chang, W.-T. (1997) Purification and characterization of two bifunctional chitinases/lysozymes extracellularly produced by Pseudomonas aeruginosa K-187 in a shrimp and crab shell powder medium, *Applied environmental microbiology.* **63**, 380-386.
260. Fukamizo, T. (2000) Chitinolytic enzymes catalysis, substrate binding, and their application, *Current protein and peptide science.* **1**, 105-124.
261. Ohno, T., Armand, S., Hata, T., Nikaidou, N., Henrissat, B., Mitsutomi, M. & Watanabe, T. (1996) A modular family 19 chitinase found in the prokaryotic organism Streptomyces griseus HUT 6037, *Journal of bacteriology.* **178**, 5065-5070.
262. Iseli, B., Boller, T. & Neuhaus, J.-M. (1993) The N-terminal cysteine-rich domain of tobacco class I chitinase is essential for chitin binding but not for catalytic or antifungal activity, *Plant physiology.* **103**, 221-226.
263. Takashima, T., Numata, T., Taira, T., Fukamizo, T. & Ohnuma, T. (2018) Structure and Enzymatic Properties of a Two-Domain Family GH19 Chitinase from Japanese Cedar (Cryptomeria japonica) Pollen, *Journal of agricultural and food chemistry.* **66**, 5699-5706.
264. Verburg, J. G. & Huynh, Q. K. (1991) Purification and characterization of an antifungal chitinase from Arabidopsis thaliana, *Plant physiology.* **95**, 450-455.
265. Datta, K., Tu, J., Oliva, N., Ona, I., Velazhahan, R., Mew, T. W., Muthukrishnan, S. & Datta, S. K. (2001) Enhanced resistance to sheath blight by constitutive expression of infection-related rice chitinase in transgenic elite indica rice cultivars, *Plant science.* **160**, 405-414.
266. Kim, J.-K., Jang, I.-C., Wu, R., Zuo, W.-N., Boston, R. S., Lee, Y.-H., Ahn, I.-P. & Nahm, B. H. (2003) Co-expression of a modified maize ribosome-inactivating protein and a rice basic chitinase gene in transgenic rice plants confers enhanced resistance to sheath blight, *Transgenic research.* **12**, 475-484.

267. Xiao, Y.-H., Li, X.-B., Yang, X.-Y., Luo, M., Hou, L., Guo, S.-H., Luo, X.-Y. & Pei, Y. (2007) Cloning and characterization of a balsam pear class I chitinase gene (Mcchit1) and its ectopic expression enhances fungal resistance in transgenic plants, *Bioscience, biotechnology, and biochemistry.* **71**, 1211-1219.
268. Liu, J.-J., Ekramoddoullah, A. K. & Zamani, A. (2005) A class IV chitinase is up-regulated by fungal infection and abiotic stresses and associated with slow-canker-growth resistance to Cronartium ribicola in western white pine (Pinus monticola), *Phytopathology.* **95**, 284-291.
269. Grover, A. (2012) Plant chitinases: genetic diversity and physiological roles, *Critical reviews in plant sciences.* **31**, 57-73.
270. Ohnuma, T., Fukuda, T., Dozen, S., Honda, Y., Kitaoka, M. & Fukamizo, T. (2012) A glycosynthase derived from an inverting GH19 chitinase from the moss Bryum coronatum, *Biochemical journal.* **444**, 437-443.
271. Ohnuma, T., Dozen, S., Honda, Y., Kitaoka, M. & Fukamizo, T. (2016) A glycosynthase derived from an inverting chitinase with an extended binding cleft, *The journal of biochemistry.* **160**, 93-100.
272. Harata, K., Schubert, W.-D. & Muraki, M. (2001) Structure of Urtica dioica agglutinin isolectin I: dimer formation mediated by two zinc ions bound at the sugar-binding site, *Acta crystallographica section D: biological crystallography.* **57**, 1513-1517.
273. Mc Grath, S. & van Sinderen, D. (2007) *Bacteriophage: genetics and molecular biology*, Caister Academic Press.
274. Kojima, M., Yoshikawa, T., Ueda, M., Nonomura, T., Matsuda, Y., Toyoda, H., Miyatake, K., Arai, M. & Fukamizo, T. (2005) Family 19 chitinase from Aeromonas sp. No. 10S-24: role of chitin-binding domain in the enzymatic activity, *Journal of biochemistry.* **137**, 235-242.
275. Ubhayasekera, W., Tang, C. M., Ho, S. W., Berglund, G., Bergfors, T., Chye, M. L. & Mowbray, S. L. (2007) Crystal structures of a family 19 chitinase from Brassica juncea show flexibility of binding cleft loops, *The FEBS journal.* **274**, 3695-3703.
276. Fujimoto, Z., Kuno, A., Kaneko, S., Kobayashi, H., Kusakabe, I. & Mizuno, H. (2002) Crystal structures of the sugar complexes of Streptomyces olivaceoviridis E-86 xylanase: sugar binding structure of the family 13 carbohydrate binding module, *Journal of molecular biology.* **316**, 65-78.
277. Coordinators, N. R. (2017) Database resources of the national center for biotechnology information, *Nucleic acids research.* **45**, D12.
278. Berman, H., Henrick, K., & Nakamura, H. (2003). Announcing the worldwide protein data bank. *Nature structural & molecular biology*. **10**, 980.
279. Consortium, U. (2007) The universal protein resource (UniProt), *Nucleic acids research.* **36**, D190-D195.
280. Edgar, R. C. (2010) Search and clustering orders of magnitude faster than BLAST, *Bioinformatics.* **26**, 2460-2461.
281. Rice, P., Longden, I. & Bleasby, A. (2000) EMBOSS: the European molecular biology open software suite, *Trends in genetics: TIG*. **16**, 276-277.
282. Henikoff, S. & Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks, *Proceedings of the national academy of sciences.* **89**, 10915-10919.
283. Tange, O. (2011) Gnu parallel-the command-line power tool, *The USENIX Magazine.* **36**, 42-47.

284. Buchholz, P. C., Zeil, C. & Pleiss, J. (2018) The scale-free nature of protein sequence space, *PloS one.* **13**, e0200815.
285. Buchholz, P. C., Fademrecht, S. & Pleiss, J. (2017) Percolation in protein sequence space, *PloS one.* **12**, e0189646.
286. Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L. & Mistry, J. (2013) Pfam: the protein families database, *Nucleic acids research.* **42**, D222-D230.
287. Eddy, S. R. (2011) Accelerated profile HMM searches, *PLoS computational biology.* **7**, e1002195.
288. Vogel, C., Widmann, M., Pohl, M. & Pleiss, J. (2012) A standard numbering scheme for thiamine diphosphate-dependent decarboxylases, *BMC biochemistry.* **13**, 24.
289. Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H. & Ferrin, T. E. (2018) UCSF ChimeraX: Meeting modern challenges in visualization and analysis, *Protein science.* **27**, 14-25.
290. Katoh, K., Kuma, K.-i., Toh, H. & Miyata, T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment, *Nucleic acids research.* **33**, 511-518.
291. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data, *Bioinformatics.* **28**, 3150-3152.
292. Verburg, J. G., Smith, C., Lisek, C. & Huynh, Q. K. (1992) Identification of an essential tyrosine residue in the catalytic site of a chitinase isolated from Zea mays that is selectively modified during inactivation with 1-ethyl-3-(3-dimethylaminopropyl)-carbodiimide, *Journal of biological chemistry.* **267**, 3886-3893.
293. Volpicella, M., Leoni, C., Fanizza, I., Distaso, M., Leoni, G., Farioli, L., Naumann, T., Pastorello, E. & Ceci, L. (2017) Characterization of maize chitinase-A, a tough allergenic molecule, *Allergy.* **72**, 1423-1429.
294. Itoh, Y., Takahashi, K., Takizawa, H., Nikaidou, N., Tanaka, H., Nishihashi, H., Watanabe, T. & Nishizawa, Y. (2003) Family 19 chitinase of Streptomyces griseus HUT6037 increases plant resistance to the fungal disease, *Bioscience, biotechnology, and biochemistry.* **67**, 847-855.
295. Kezuka, Y., Ohishi, M., Itoh, Y., Watanabe, J., Mitsutomi, M., Watanabe, T. & Nonaka, T. (2006) Structural studies of a two-domain chitinase from Streptomyces griseus HUT6037, *Journal of molecular biology.* **358**, 472-484.
296. Akagi, K.-i., Watanabe, J., Hara, M., Kezuka, Y., Chikaishi, E., Yamaguchi, T., Akutsu, H., Nonaka, T., Watanabe, T. & Ikegami, T. (2006) Identification of the substrate interaction region of the chitin-binding domain of Streptomyces griseus chitinase C, *Journal of biochemistry.* **139**, 483-493.
297. Tang, C. M., Chye, M.-L., Ramalingam, S., Ouyang, S.-W., Zhao, K.-J., Ubhayasekera, W. & Mowbray, S. L. (2004) Functional analyses of the chitin-binding domains and the catalytic domain of Brassica juncea chitinase BjCHI1, *Plant molecular biology.* **56**, 285-298.
298. Yamagami, T. & Funatsu, G. (1993) Purification and some properties of three chitinases from the seeds of rye (Secale cereale), *Bioscience, biotechnology, and biochemistry.* **57**, 643-647.

299. Yamagami, T. & Funatsu, G. (1995) Identification of the tryptophan residue located at the substrate-binding site of rye seed chitinase-c, *Bioscience, biotechnology, and biochemistry.* **59**, 1076-1081.
300. Ohnuma, T., YAGi, M., Yamagami, T., Taira, T., Aso, Y. & Ishiguro, M. (2002) Molecular cloning, functional expression, and mutagenesis of cDNA encoding rye (Secale cereale) seed chitinase-c, *Bioscience, biotechnology, and biochemistry.* **66**, 277-284.
301. Hoster, F., Schmitz, J. E. & Daniel, R. (2005) Enrichment of chitinolytic microorganisms: isolation and characterization of a chitinase exhibiting antifungal activity against phytopathogenic fungi from a novel Streptomyces strain, *Applied microbiology and biotechnology.* **66**, 434-442.
302. Osswald, W. F., Shapiro, J. P., Doostdar, H., McDonald, R. E., Niedz, R. P., Nairn, C. J., Hearn, C. J. & Mayer, R. T. (1994) Identification and characterization of acidic hydrolases with chitinase and chitosanase activities from sweet orange callus tissue, *Plant and cell physiology.* **35**, 811-820.
303. Nielsen, K. K., Bojsen, K., Roepstorff, P. & Mikkelsen, J. D. (1994) A hydroxyproline-containing class IV chitinase of sugar beet is glycosylated with xylose, *Plant molecular biology.* **25**, 241-257.
304. Schultze, M., Staehelin, C., Brunner, F., Genetet, I., Legrand, M., Fritig, B., Kondorosi, E. & Kondorosi, Á. (1998) Plant chitinase/lysozyme isoforms show distinct substrate specificity and cleavage site preference towards lipochitooligosaccharide Nod signals, *The plant journal.* **16**, 571-580.
305. Yerzhebayeva, R., Abekova, A., Konysbekov, K., Bastaubayeva, S., Kabdrakhmanova, A., Absattarova, A. & Shavrukov, Y. (2018) Two sugar beet chitinase genes, BvSP2 and BvSE2, analysed with SNP Amplifluor-like markers, are highly expressed after Fusarium root rot inoculations and field susceptibility trial, *PeerJ.* **6**, e5127.
306. Takakura, Y., Ito, T., Saito, H., Inoue, T., Komari, T. & Kuwata, S. (2000) Flower-predominant expression of a gene encoding a novel class I chitinase in rice (Oryza sativa L.), *Plant molecular biology.* **42**, 883-897.
307. Shinshi, H., Mohnen, D. & Meins, F. (1987) Regulation of a plant pathogenesis-related enzyme: inhibition of chitinase and chitinase mRNA accumulation in cultured tobacco tissues by auxin and cytokinin, *Proceedings of the national academy of sciences.* **84**, 89-93.
308. Yeh, S., Moffatt, B. A., Griffith, M., Xiong, F., Yang, D. S., Wiseman, S. B., Sarhan, F., Danyluk, J., Xue, Y. Q. & Hew, C. L. (2000) Chitinase genes responsive to cold encode antifreeze proteins in winter cereals, *Plant physiology.* **124**, 1251-1264.
309. Huet, J., Wyckmans, J., Wintjens, R., Boussard, P., Raussens, V., Vandenbussche, G., Ruysschaert, J. M., Azarkan, M. & Looze, Y. (2006) Structural characterization of two papaya chitinases, a family GH19 of glycosyl hydrolases, *Cellular and molecular life sciences CMLS.* **63**, 3042-3054.
310. Song, H. K. & Suh, S. W. (1996) Refined structure of the chitinase from barley seeds at 2.0 Å resolution, *Acta crystallographica section D: biological crystallography.* **52**, 289-298.
311. ANDERSEN, M. D., JENSEN, A., ROBERTUS, J. D., Robert, L. & SKRIVER, K. (1997) Heterologous expression and characterization of wild-type and mutant forms of a 26 kDa endochitinase from barley (Hordeum vulgare L.), *Biochemical journal.* **322**, 815-822.

312. Hollis, T., Honda, Y., Fukamizo, T., Marcotte, E., Day, P. J. & Robertus, J. D. (1997) Kinetic analysis of barley chitinase, *Archives of biochemistry and biophysics.* **344**, 335-342.
313. Ohnishi, T., Juffer, A. H., Tamoi, M., Skriver, K. & Fukamizo, T. (2005) 26 kDa endochitinase from barley seeds: an interaction of the ionizable side chains essential for catalysis, *Journal of biochemistry.* **138**, 553-562.
314. Mizuno, R., Itoh, Y., Nishizawa, Y., Kezuka, Y., Suzuki, K., Nonaka, T. & Watanabe, T. (2008) Purification and characterization of a rice class I chitinase, OsChia1b, produced in Esherichia coli, *Bioscience, biotechnology, and biochemistry.* **72**, 893-895.
315. Kezuka, Y., Kojima, M., Mizuno, R., Suzuki, K., Watanabe, T. & Nonaka, T. (2010) Structure of full-length class I chitinase from rice revealed by X-ray crystallography and small-angle X-ray scattering, *Proteins: structure, function, and bioinformatics.* **78**, 2295-2305.
316. Kaomek, M., Mizuno, K., Fujimura, T., Sriyotha, P. & Cairns, J. R. K. (2003) Cloning, expression, and characterization of an antifungal chitinase from Leucaena leucocephala de Wit, *Bioscience, biotechnology, and biochemistry.* **67**, 667-676.
317. Xu, Y., Zhu, Q., Panbangred, W., Shirasu, K. & Lamb, C. (1996) Regulation, expression and function of a new basic chitinase gene in rice (Oryza sativa L.), *Plant molecular biology.* **30**, 387-401.
318. Nielsen, K., Jørgensen, P. & Mikkelsen, J. (1994) Antifungal activity of sugar beet chitinase against Cercospora beticola: an autoradiographic study on cell wall degradation, *Plant pathology.* **43**, 979-986.
319. Mavrodi, D. V., Loper, J. E., Paulsen, I. T. & Thomashow, L. S. (2009) Mobile genetic elements in the genome of the beneficial rhizobacterium Pseudomonas fluorescens Pf-5, *BMC microbiology.* **9**, 8.
320. Chen, A., Yu, L., Fan, J., Feng, D. & Wang, J. (2008) The expression, purification and activity analysis of the rice chitinase gene in Escherichia coli, *Sheng wu gong cheng xue bao= Chinese journal of biotechnology.* **24**, 188-192.
321. Lerner, D. R. & Raikhel, N. V. (1992) The gene for stinging nettle lectin (Urtica dioica agglutinin) encodes both a lectin and a chitinase, *Journal of biological chemistry.* **267**, 11085-11091.
322. Does, M. P., Houterman, P. M., Dekker, H. L. & Cornelissen, B. J. (1999) Processing, targeting, and antifungal activity of stinging nettle agglutinin in transgenic tobacco, *Plant physiology.* **120**, 421-432.
323. Paszota, P., Escalante-Perez, M., Thomsen, L. R., Risør, M. W., Dembski, A., Sanglas, L., Nielsen, T. A., Karring, H., Thøgersen, I. B. & Hedrich, R. (2014) Secreted major Venus flytrap chitinase enables digestion of arthropod prey, *Biochimica et biophysica acta (BBA)-proteins and proteomics.* **1844**, 374-383.
324. Kragh, K. M., Hendriks, T., de Jong, A. J., Schiavo, F. L., Bucherna, N., Højrup, P., Mikkelsen, J. D. & de Vries, S. C. (1996) Characterzation of chitinases able to rescue somatic embryos of the temperature-sensitive carrot variantts11, *Plant molecular biology.* **31**, 631-645.
325. Kolosova, N., Breuil, C. & Bohlmann, J. (2014) Cloning and characterization of chitinases from interior spruce and lodgepole pine, *Phytochemistry.* **101**, 32-39.
326. Tsujibo, H., Okamoto, T., Hatano, N., Miyamoto, K., Watanabe, T., Mitsutomi, M. & Inamori, Y. (2000) Family 19 chitinases from Streptomyces thermoviolaceus OPC-520:

molecular cloning and characterization, *Bioscience, biotechnology, and biochemistry.* **64**, 2445-2453.
327. Yano, S., Rattanakit, N., Wakayama, M. & Tachiki, T. (2004) A chitinase indispensable for formation of protoplast of Schizophyllum commune in basidiomycete-lytic enzyme preparation produced by Bacillus circulans KA-304, *Bioscience, biotechnology, and biochemistry.* **68**, 1299-1305.
328. Yano, S., Rattanakit, N., Wakayama, M. & TACHIKI, T. (2005) Cloning and expression of a Bacillus circulans KA-304 gene encoding chitinase I, which participates in protoplast formation of Schizophyllum commune, *Bioscience, biotechnology, and biochemistry.* **69**, 602-609.
329. Yano, S., Suyotha, W., Honda, A., Takagi, K., Rattanakit-Chandet, N., Wakayama, M. & Tachiki, T. (2011) N-terminal region of chitinase I of Bacillus circulans KA-304 contained new chitin-biding domain, *Bioscience, biotechnology, and biochemistry.* **75**, 299-304.
330. Saito, A., Miyashita, K., Biuković, G. & Schrempf, H. (2001) Characteristics of a Streptomyces coelicolor A3 (2) extracellular protein targeting chitin and chitosan, *Applied environmental microbiology.* **67**, 1268-1273.
331. Heggset, E. B., Hoell, I. A., Kristoffersen, M., Eijsink, V. G. & Vårum, K. M. (2009) Degradation of chitosans with chitinase G from Streptomyces coelicolor A3 (2): production of chito-oligosaccharides and insight into subsite specificities, *Biomacromolecules.* **10**, 892-899.
332. Nakamura, T., Ishikawa, M., Nakatani, H. & Oda, A. (2008) Characterization of cold-responsive extracellular chitinase in bromegrass cell cultures and its relationship to antifreeze activity, *Plant physiology.* **147**, 391-401.
333. Allona, I., Collada, C., Casado, R., Paz-Ares, J. & Aragoncillo, C. (1996) Bacterial expression of an active class Ib chitinase from Castanea sativa cotyledons, *Plant molecular biology.* **32**, 1171-1176.
334. Garcia-Casado, G., Collada, C., Allona, I., Casado, R., Pacios, L. F., Aragoncillo, C. & Gomez, L. (1998) Site-directed mutagenesis of active site residues in a class I endochitinase from chestnut seeds, *Glycobiology.* **8**, 1021-1028.
335. Schlesier, B., Koch, G. & Horstmann, C. (1998) Characterization of a class II chitinase from jack bean (Canavalia ensiformis) seeds, *Food/Nahrung.* **42**, 170-170.
336. Hahn, M., Hennig, M., Schlesier, B. & Höhne, W. (2000) Structure of jack bean chitinase, *Acta crystallographica section D: biological crystallography.* **56**, 1096-1099.
337. Huynh, Q. K., Hironaka, C. M., Levine, E. B., Smith, C., Borgmeyer, J. & Shah, D. (1992) Antifungal proteins from plants. Purification, molecular cloning, and antifungal properties of chitinases from maize seed, *Journal of biological chemistry.* **267**, 6635-6640.
338. Liu, Z. H., Wang, Y. C., Qi, X. T. & Yang, C. P. (2010) Cloning and characterization of a chitinase gene Lbchi31 from Limonium bicolor and identification of its biological activity, *Molecular biology reports.* **37**, 2447-2453.
339. Huang, L., Garbulewska, E., Sato, K., Kato, Y., Nogawa, M., Taguchi, G. & Shimosaka, M. (2012) Isolation of genes coding for chitin-degrading enzymes in the novel chitinolytic bacterium, Chitiniphilus shinanonensis, and characterization of a gene coding for a family 19 chitinase, *Journal of bioscience and bioengineering.* **113**, 293-299.
340. Robinson, S. P., Jacobs, A. K. & Dry, I. B. (1997) A class IV chitinase is highly expressed in grape berries during ripening, *Plant physiology.* **114**, 771-778.

341. Taira, T., Yamagami, T., Aso, Y., Ishiguro, M. & Ishihara, M. (2001) Localization, accumulation, and antifungal activity of chitinases in rye (Secale cereale) seed, *Bioscience, biotechnology, and biochemistry.* **65**, 2710-2718.
342. Ohnuma, T., Taira, T., Yamagami, T., Aso, Y. & Ishiguro, M. (2004) Molecular cloning, functional expression, and mutagenesis of cDNA encoding class I chitinase from rye (Secale cereale) seeds, *Bioscience, biotechnology, and biochemistry.* **68**, 324-332.
343. Payne, G., Ahl, P., Moyer, M., Harper, A., Beck, J., Meins, F. & Ryals, J. (1990) Isolation of complementary DNA clones encoding pathogenesis-related proteins P and Q, two acidic chitinases from tobacco, *Proceedings of the national academy of sciences.* **87**, 98-102.
344. Chlan, C. A. & Bourgeois, R. P. (2001) Class I chitinases in cotton (Gossypium hirsutum): characterization, expression and purification, *Plant science.* **161**, 143-154.
345. Wiweger, M., Farbos, I., Ingouff, M., Lagercrantz, U. & Von Arnold, S. (2003) Expression of Chia4-Pa chitinase genes during somatic and zygotic embryo development in Norway spruce (Picea abies): similarities and differences between gymnosperm and angiosperm class IV chitinases, *Journal of experimental botany.* **54**, 2691-2699.
346. Verburg, J. G., Rangwala, S. H., Samac, D. A., Luckow, V. A. & Huynh, Q. K. (1993) Examination of the role of tyrosine-174 in the catalytic mechanism of the Arabidopsis thaliana-chitinase: comparison of variant chitinases generated by site-directed mutagenesis and expressed in insect cells using baculovirus vectors, *Archives of biochemistry and biophysics.* **300**, 223-230.
347. Thomma, B. P., Eggermont, K., Penninckx, I. A., Mauch-Mani, B., Vogelsang, R., Cammue, B. P. & Broekaert, W. F. (1998) Separate jasmonate-dependent and salicylate-dependent defense-response pathways in Arabidopsis are essential for resistance to distinct microbial pathogens, *Proceedings of the national academy of sciences.* **95**, 15107-15111.
348. Wemmer, T., Kaufmann, H., Kirch, H.-H., Schneider, K., Lottspeich, F. & Thompson, R. D. (1994) The most abundant soluble basic protein of the stylar transmitting tract in potato (Solanum tuberosum L.) is an endochitinase, *Planta.* **194**, 264-273.
349. O'Riordain, G., Radauer, C., Hoffmann-Sommergruber, K., Adhami, F., Peterbauer, C., Blanco, C., Godnic-Cvar, J., Scheiner, O., Ebner, C. & Breiteneder, H. (2002) Cloning and molecular characterization of the Hevea brasiliensis allergen Hev b 11, a class I chitinase, *Clinical & experimental allergy.* **32**, 455-462.
350. Yano, S., Rattanakit, N., Honda, A., Noda, Y., Wakayama, M., Plikomol, A. & Tachiki, T. (2008) Purification and characterization of chitinase A of Streptomyces cyaneus SP-27: an enzyme participates in protoplast formation from Schizophyllum commune mycelia, *Bioscience, biotechnology, and biochemistry.* **72**, 54-61.
351. Yano, S., Honda, A., Rattanakit-Chandet, N., Noda, Y., Wakayama, M., Plikomol, A. & Tachiki, T. (2009) Role of chitin binding domain of chitinase A of streptomyces cyaneus SP-27 in protoplast formation from schizophyllum commune, *Bioscience, biotechnology, and biochemistry.* **73**, 733-735.
352. Sticher, L., Hofsteenge, J., Neuhaus, J.-M., Boller, T. & Meins Jr, F. (1993) Posttranslational Processing of a New Class of Hydroxyproline-Containing Proteins (Prolyl Hydroxylation and C-Terminal Cleavage of Tobacco (Nicotiana tabacum) Vacuolar Chitinase), *Plant physiology.* **101**, 1239-1247.

353. Freydl, E., Meins, F., Boller, T. & Neuhaus, J.-M. (1995) Kinetics of prolyl hydroxylation, intracellular transport and C-terminal processing of the tobacco vacuolar chitinase, *Planta.* **197**, 250-256.
354. Iseli-Gamboni, B., Boller, T. & Neuhaus, J.-M. (1998) Mutation of either of two essential glutamates converts the catalytic domain of tobacco class I chitinase into a chitin-binding lectin, *Plant science.* **134**, 45-51.
355. Harikrishna, K., Jampates-Beale, R., Milligan, S. B. & Gasser, C. S. (1996) An endochitinase gene expressed at high levels in the stylar transmitting tissue of tomatoes, *Plant molecular biology.* **30**, 899-911.
356. Tsujibo, H., Kubota, T., Yamamoto, M., Miyamoto, K. & Inamori, Y. (2003) Characterization of chitinase genes from an alkaliphilic actinomycete, Nocardiopsis prasina OPC-131, *Applied environmental microbiology.* **69**, 894-900.
357. Fujimura, T., Shigeta, S., Suwa, T., Kawamoto, S., Aki, T., Masubuchi, M., Hayashi, T., Hide, M. & Ono, K. (2005) Molecular cloning of a class IV chitinase allergen from Japanese cedar (Cryptomeria japonica) pollen and competitive inhibition of its immunoglobulin E-binding capacity by latex C-serum, *Clinical & experimental allergy.* **35**, 234-243.
358. Takashima, T., Ohnuma, T. & Fukamizo, T. (2017) NMR assignments and ligand-binding studies on a two-domain family GH19 chitinase allergen from Japanese cedar (Cryptomeria japonica) pollen, *Biomolecular NMR assignments.* **11**, 85-90.
359. Nakayama, K., Takashima, K., Ishihara, H., Shinomiya, T., Kageyama, M., Kanaya, S., Ohnishi, M., Murata, T., Mori, H. & Hayashi, T. (2000) The R-type pyocin of Pseudomonas aeruginosa is related to P2 phage, and the F-type is related to lambda phage, *Molecular microbiology.* **38**, 213-231.