Rendiconti Lincei. Scienze Fisiche e Naturali (2022) 33:449–460
https://doi.org/10.1007/s12210-022-01068-7

RESEARCH PAPER

Phylogenetic study to analyse the evolutionary relationship of taxonomically diverse α-amylases

Sachin Kajla[1] · Ritu Kumari[1] · Anima Das[1] · Vikas Kumar Patel[1]

Received: 25 October 2021 / Accepted: 19 March 2022 / Published online: 1 April 2022

Abstract
Microbial α-amylases have been identified and characterized to understand their progressive role in industrial applications. Computational biology tools are employed in various *in-silico* studies for finding out the ways to improve the titre of this enzyme with improved stability and specificity. This study includes detailed evolutionary analysis and comparison of α-amylase amino acid sequences found in a wide spectrum of microorganisms like algae, bacteria, fungi, cyanobacteria, and plants. We analysed the conservation patterns and signature modules in molecular architectures to study the phylogenetic relationship among various microbial and plant taxa. Additionally, efforts were made to identify signal peptides responsible for the secretion of α-amylase protein outside the cell. Our results concluded the presence of different types of domains with their single or multiple copies, and distinct signal peptides which may alter the overall functioning and secretion efficiency of enzyme. Collectively, overall molecular diversity leads to adaptive evolution of α-amylase in different microbes and plants.

Keywords α-amylase · Evolutionary analysis · Conserved domains · Signal peptides · Adaptive evolution

1 Introduction

Enzymes are biocatalysts that can enhance the reaction rate by decreasing the free energy of the transition phase (Choudhury 2020; Robinson 2015). Enzymes are highly specific to uniquely recognize their substrate which makes them suitable for sophisticated biological reactions (Bhatia 2018). Enzymes play a crucial role in various industries; they are used in production of diverse valuable products. Amylases are one of the most explored among the industrially important enzymes; and considered for various kinds of applications in food, animal feed, detergents, textile, pharmaceutical, and brewing industries (Hmidet et al. 2009; Tiwari et al. 2015). Amino acid sequences for α-amylases can be identified from a diverse group of organisms like plants, bacteria, and fungi. However, their functionality and efficiency depend on the molecular architecture of enzyme. Microbial α-amylases are extremely valuable because they are extracellular in nature, relatively easier to produce them at large-scale and recover them from the fermentation tanks with minimal economical inputs (Mukherjee et al. 2009; Paul 2016).

Global industrial production of α-amylases from microbes comprises up to 30% of the enzyme's market (Balakrishnan et al. 2021). According to the grand view research report, global baking α-amylase market is projected to reach around USD 320.1 million by 2024 (https://www.grandviewresearch.com/press-release/global-alpha-amylase-baking-enzyme-market). Ethanol and fructose syrups constitute up to 72% and 40% of the sugar yields, respectively (http://www.starch.dk/isi/market/index.asp). α-amylase plays significant role in different industries such as bakery, textile, paper, and high fructose corn syrups (HFCS). In textile industry, starch is hydrolysed and solubilized using α-amylases to provide stiffness to the clothes (Mehta and Satyanarayana 2016; Tiwari et al. 2015). Moreover, α-amylases play important role in sugar and glucose industry due to their ability to

Sachin Kajla and Vikas Kumar Patel both corresponding authors contributed equally.

Sachin Kajla and Ritu Kumari both authors contributed equally to this work.

✉ Sachin Kajla
skajla09@gmail.com

✉ Vikas Kumar Patel
vikaspatel15@gmail.com

[1] Environmental Research Group, R & D and Scientific Services Department, Tata Steel Limited, Jamshedpur 831007, India

break polysaccharides in monomers (Nielsen and Borchert 2000). Ethyl alcohol is also manufactured from the conversion of starch present in grain and potatoes with the activity of α-amylases and subsequently converted to ethanol via co-cultivation of yeast (Kumar and Singh 2016). They also play important role in biofuel production (Kumar and Singh 2016).

Amylases belong to the class of hydrolases and are of different kinds with diverse mechanism of action. Amylase can be divided into two groups; endo-amylase and exo-amylase. Endo-amylase includes α-amylase (α-1,4glucan-glucanohydrolase, EC 3.2.1.1) that hydrolyses α-1,4-glycosidic bonds of starch randomly at the interior sites and forms oligosaccharides of various lengths. Exo-amylase includes the β-amylase and γ-amylase which hydrolyses starch at non-reducing ends. *α*-amylases are metalloenzymes that depend on the calcium ions for their activity (de Souza Vandenberghe et al. 2020; El-Enshasy et al. 2013). Cyclodextrins are hydrolysed from a different category of enzymes i.e., catalytically flexible carbohydrases. These are known as maltogenic amylases (EC 3.2.1.133) which uniquely possess additional 130 residues at N-terminus that are altogether absent in natural amylases. These are also capable of hydrolysing α-1,4 and α-1,6 glycosidic linkages present in substrates along with trans-glycosylation of hydrolytic substances (Li et al. 2011). β-amylases (α-1,4-glucan maltohydrolase, EC 3.2.1.2) act as an exoenzyme that hydrolyses α-1,4-glycosidic bonds of starch molecules and produce β-maltose from non-reducing ends. It also cleaves α-1,4 and α-1,6 glycosidic bonds like amyloglucosidase and α-glucosidase. Plant and bacterial β-amylases have some variations; bacterial amylases are capable of binding and hydrolysing raw starch which is not observed in plant amylases (Martin et al. 2019). α-glucosidase (E.C.3.2.1.20), starch hydrolysing enzyme, releases D-glucose units by acting upon α- 1,4 and α- 1,6-oligosaccharide linkages from non-reducing ends. γ-amylase (Glucan 1,4-α-glucosidase, EC 3.2.1.3) cleaves the terminal α-1,4-glycosidic linkages to partially hydrolyse the starch after liquefaction. Glucose is produced as a product after hydrolysis of starch with γ-amylase (Mehta and Satyanarayana 2016; Tiwari et al. 2015).

α-amylase consists of single polypeptide chain with three domains 'A', 'B' and 'C'. The 'A' domain is the most conserved domain consisting of eight parallel β-sheets surrounded by eight α-helices $(\beta/\alpha)_8$ (Farber and Petsko 1990). The 'B' domain is entangled in between the 'A' and 'C' domains and attached to 'A' domain with disulphide bonds. The 'C' domains are quite conserved and consist of antiparallel β-barrel with unknown function (Bayer et al. 1995). The active site of the α-amylase resides in the cleft of domains 'A' and 'B'. The binding sites of Ca^{+} ions reside in the interface of 'A' and 'B' domains. These ions also help in allosteric activation and stabilization of the three-dimensional structure of α-amylases (Linden et al. 2003; Muralikrishna and Nirmala 2005).

Among the amylases; α-amylases have been studied in-depth, a lot of computational research has been done; however, the probable cause behind their differential functioning has not been studied yet in detail. α-amylases are the foremost amylases and are used extensively in various industrial applications. They have been studied immensely due to their structural and functional aspects. Three-dimensional structure of α-amylase revealed about 30% sequence identity from different organisms and they chiefly belong to the family 13 of Glycosyl Hydrolase (GH13) (Henrissat and Bairoch 1993). According to the CAZy database, GHs of α-amylases are categorized into four-distinct families depending on their sequence analysis i.e., GH13, GH 57, GH119, and GH126, respectively. GH13 is the predominant family consisting of 124,300 sequences, chiefly occurring in microbes including eukaryotes, bacteria, and archaea (Janeček et al. 2014; Lombard et al. 2014). On the other hand, GH57 family is majorly present in archaea, bacteria, and some of the eukaryotes. GH119 and GH126 could be predicted in bacterial species only (Sidar et al. 2020).

In general, GH13 α-amylase is a polyspecific enzyme that contains a $(\beta/\alpha)_8$ or TIM barrel structure constituting catalytic site residues. The stability of TIM barrel is maintained by seven highly conserved regions (CSRs) present in primary sequences which consist of catalytic and essential amino acids. According to CAZy database, GH 13 is associated with the GH-H clan comprising of GH70 and GH77 families at higher levels and in a lower hierarchy, it is subdivided into 44 curator designed GH13 subfamilies which are involved in hydrolysis, isomerization, and transglycosylation mechanisms (Janeček and Gabriško 2016; Janeček and Svensson 2022; Janeček et al. 2014; Stam et al. 2006). GH 57 α-amylase family is the second family containing ~ 900 representatives and each member is identified by five CSRs that are distinct from the GH13 family (Janeček et al. 2014). The prime reason for classifying GH57 family is identification of two distinct amino acid sequences from two microorganisms i.e. *Dictyoglomus thermophilum* (bacterium) (FUKUSUMI et al. 1988) and *Pyrococcus furiosus* (archaea) (Laderman et al. 1993). GH57 family consists of an incomplete TIM barrel or $(\beta/\alpha)_7$ which have fold bearing catalytic residues where aspartic acid acts as a proton donor and glutamic acid as a nucleophile (Janeček and Svensson 2022). In CAZy database, GH119 is considered as the third and smallest GH α-amylase family containing only 38 bacterial representatives. α-amylase IgtZ reported in *Bacillus circulans* is still the individual represented member of GH119 family which acts on maltotetraose, soluble starch, yielding glucose and maltooligosaccharides (Watanabe et al. 2006). Based on an in-silico study of 2012, the structure of GH119

was predicted to have a partial TIM barrel in the catalytic domain which is closely related to GH57 family (Janeček et al. 2014). The fourth α-amylase family is GH126, reported in 2011, in three- dimensional architecture of CPF_2247 protein found in genome of *Clostridium perfringens* (Ficko-Blean et al. 2011). This structure consists of $(\alpha/\alpha)_{\text{sixfold}}$ barrel which differs completely from the partial TIM barrels found in GH13 and GH57 family members (Janeček and Svensson 2022).

Starch binding domains (SBDs) present in different amino acid sequences of α-amylases are grouped under carbohydrate-binding modules (CBMs). SBDs are further categorized into 67 protein families based on sequence similarity of amino acids and alterations present in the ligand specificity. CBMs are recognized as a large association of protein domains with no catalytic function. These are generally associated with GH enzymes and act as substrate-binding modules by directing the insoluble substrate towards enzymes, consequently, enhancing the hydrolysis. It is proposed that the catalytic function of CAZymes can be enhanced by the binding ability of CBM; which directs the enzyme towards the substrate and increases the enzyme–substrate interaction efficiency (Boraston et al. 2002; Coutinho 1999; Gangadharan et al. 2020). Hence, the elimination of CBMs from enzymes may further lead to minimized enzymatic action with diminished stability (Bernardes et al. 2019; Cockburn et al. 2018). Starch hydrolysing enzymes found in CBM are termed "SBDs" which is a continuous sequence of the polypeptide chain. Polysaccharides are degraded by carbohydrate active enzymes which bind to the carbohydrate-binding sites present at a distance from the active site of an enzyme. Additionally, these carbohydrate-binding sites can be present in CBMs or surface-binding sites (SBSs). The presence of SBD with a specific module of a protein may not be significantly related to the binding of raw starch to amylolytic enzyme, since; the accumulation of extra aromatic residues forming SBSs in the enzyme surface could be a possible reason behind this (Janeček et al. 2019; Nielsen et al. 2009). SBDs are classified mainly among the seven families namely CBM20, CBM21, CBM 25, CBM26, CBM34, CBM41, and CBM48. In most of the α-amylases, SBDs primarily belong to CBM 20 and CBM 25 families (Gangadharan et al. 2020; Mehta and Satyanarayana 2016; Sidar et al. 2020). A detailed study about the structure, function, and evolution of various starch binding domains as CBM modules have been discussed in a recent study (Janeček et al. 2019).

Molecular evolution among diverse taxa can be inferred by analysing the inconsistencies in amino acid sequence, mapping their conserved regions, and analysing their pattern of evolution through phylogenetic studies. It helps to understand the genetic relationship between diverse taxa. Evolutionary classification with computational biology tools is helpful to predict the ancestral history and genetic relationship of coding amino acid sequences of a functional protein from diverse taxa with the help of phylogenetic tree, cladogram and dendrogram (Choudhuri 2014; Podani and Morrison 2017).

As discussed above, α-amylases are vital in functions for numerous industries; these enzymes usually show variations in their catalytic activities among phylogenetically diverse taxa. In the present study, we analysed the phylogenetically diverse α-amylases from different taxonomic groups and tried to find out the variations in conserved domains and active sites of α-amylases amino acid sequences retrieved from diverse taxa. Inferences were made to find out the dissimilarity of amino acids in hydrolytic domains and presence of mutations or multiple copies of either mutated or original carbohydrate-binding domains contribute to adaptive radiation of α-amylases in different clades during evolution. It may be a reason behind their differential efficacies towards starch degradation.

2 Materials and methods

2.1 Sequence collection, comparison and prediction of conserved domains

The full-length non-redundant α-amylase amino acid sequences of photosynthetic and non-photosynthetic microbes and model plant *Arabidopsis thaliana* were retrieved from UniProt KB knowledge database (UniProt: the universal protein knowledgebase in 2021, 2021). Amino acid sequences were aligned using Clustal Omega program (Sievers et al. 2011) to identify the conserved regions and consensus sequences among diverse taxa of microbes and model plant *Arabidopsis* (Madeira et al. 2019). The Clustal omega aligned amino acid sequence files were further annotated using Jalview 2.11.1.4. (Waterhouse et al. 2009).

All the conserved domains present in α-amylase amino acid sequences of photosynthetic and non-photosynthetic microorganisms and higher plants were identified and analysed using Batch CD Search Tool (Lu et al. 2020); a freely accessible tool of Conserved Domain Database (CDD) governed by NCBI (National Center for Biotechnology Information Software) (Marchler-Bauer and Bryant 2004; Marchler-Bauer et al. 2010). A threshold cut-off of 0.01 was used for identifying various gene super-families. Based on the observations; conserved domains, functional sites, motifs and protein super-families were predicted and evaluated. Furthermore, Superfamily 1.75 server (http://supfam.org) was used to detect diverse and detailed family and superfamily present in various amino acid sequences of diverse α-amylases (Gough et al. 2001). The domain architecture

of 197 amino acid sequences of α-amylase was determined using Conserved Domain Architecture Retrieval Tool (CDART) (Geer et al. 2002).

2.2 Prediction of signal peptides

Signal peptide sequences along with cleavage positions were determined and analysed by PrediSi server (http://www.predisi.de/) (Hiller et al. 2004). A minimum threshold cut-off of 0.5 was used to predict the presence of peptides.

2.3 Evolutionary analysis

Multiple sequence alignment of 197 α-amylase amino acid sequences was done using Clustal W program and phylogenetic analysis was performed using Mega 10.2.4 tool (Kumar et al. 2018). The Neighbour-Joining (N-J) method was used for creation of bootstrap phylogenetic tree with 500 bootstrap replications. The results were saved in newick format from MEGA 10.2.4, which were further annotated using iTOL v5 (Interactive tree of life) (https://itol.embl.de/login.cgi) (Letunic and Bork 2021). The amino acid sequence of endoglucanase of *Bacillus subtilis* (tr-Q93LD0) was used as an outgroup member. Another tree was also made for understanding signal peptides evolution. For doing so, the amino acid sequences of various signal peptides were aligned using Clustal W, and phylogenetic tree was constructed using Mega 10.2.4 package, and finally, results were annotated by iTOL v5.

3 Results and discussion

A total of 197 α-amylase amino acid sequences including; 25 sequences from algae, 33 sequences from cyanobacteria, 46 sequences from fungi, 4 from *Arabidopsis thaliana* and 89 sequences from bacteria were used. The accession numbers of amino acid sequences with corresponding names of different taxa are listed in Supplementary Table 1. The present bioinformatics study provides detailed information about phylogenetically diverse α-amylases consisting of unique GH13 (chiefly), GH57 and GH119 protein families. These non-redundant full-length amino acid sequences had diverse amino acid compositions and showed inconsistencies in their occurrences. These sequences mainly belonged to three protein groups i.e. α-amylase, 1,4- alpha glucan branching enzyme and maltogenic amylase. The α-amylase sequences having GH57 and GH119 and GH13 were evaluated and studied the evolutionary relatedness of these sequences in diverse taxa. The alignment of various α-amylase protein sequences was performed by taking entire lengths of amino acid sequences from different taxa. The results from multiple sequence alignment showed the conservation pattern among diverse α-amylase sequences present in various microorganisms. Additionally, it also helped us to identify the internal mutations present in amino acid sequences at different positions (Supplementary Fig. 1).

3.1 α-amylase proteins present in microorganisms and higher plants belong to diverse gene families

Molecular mechanisms and divergence of *α-amylase* amino acid sequences among closely related groups of organisms can be predicted by categorizing them based on families, super-families, sequence similarities and structural framework. Moreover, this also helps in recognizing the functional alterations among organisms (Todd et al. 2001). Protein super-families are described as the combination of one or more protein families which helps in identifying the closely related species that emerged separately during evolution. From the CDD search tool, we identified unique super-families present in distinct microorganisms (Supplementary table 2). In case of photosynthetic organisms like algae; majorly three super-families including; *cl33494, cl38930, cl29240* were observed. In *Chlorella sorokiniana* (A0A2P6TQG1), three different super-families i.e., *cl38930* (AmyAc), *cl29240* (Alpha-amyl C2) and *cl02663* (Fasciclin) were found. In contrary, *Trebouxia* (A0A5J4XVP9) consisted *cl09141*. The most common superfamily found in cyanobacterial species was *cl38930*. In *Gloeothece citriformis* (B7K8V9); *cl38930* was present along with additional CBM 20 (*cl15347* type) and AmyAc superfamily. *Cyanothece* sp. (A0A3B8YBI3) consists of AmyAc, CBM20 and Malt amylase C superfamily. In case of fungal sequences; *cl38930* and *cl07771* were most commonly found super-families. The fungi *Rhizoctonia solani* had three distinct super-families including; *cl38930* (AmyAc), *cl15347* (CBM20) and *cl02706* (Aamy C). In *Ophiostoma floccosum* and *Emericella nidulans*, the presence of *cl38930, cl07771* and *cl15347* (CBM 20 glucoamylase) was identified. In higher plants, like *Arabidopsis thaliana*, three conserved domains reported were from the *cl38930, cl33494* and *cl33565* families. It was predicted that in most of the bacterial species which were included in this bioinformatics study consisted of conserved domains such as family *cl38930* and *cl2706*. An additional alpha-amylase N domain (*cl38100*) was found in *Bacillus licheniformis* (Q04977) whereas, in *Bacillus amyloliquefaciens* and *Bacillus subtilis* conserved carbohydrate-binding motif; CBM 26 (*cl23798*) was detected. In a member of GH57 family, *Geobacillus stearothermophilus*, the presence of conserved malt amylase C domain was observed with *cl38930* superfamily. Six distinct domains were found in *Bacillus circulans* (sp-A0P8X0) that consisted of hydrolytic domain GH119 (*cl15692* type), two copies of CBM 25 family (*cl23798* type), additional CBM2

(*cl15347* type), unique FN3 (*cl21522* type) and TALPID3 superfamily, respectively. A unique type of superfamily *cl30295* was detected in archaea; *Saccharolobus solfataricus*. Out of eight different α-amylase sequences of distinct taxa present in *Streptomyces;* CBM 20 domain was present only in six amino acid sequences. However, additional multiple domains like AamyC and AmyAc were present in all eight sequences. Detailed information about the families and super-families were extracted from Superfamily 1.75 online server which predicted different types of protein super-families in α-amylase sequences of diverse taxa (Supplementary Table 3). Functional diversity in super-families is due to variations present in sequences and domains. Also, substrate specificity and position of catalytic amino acid residues may differ among distinct groups of organisms (Gough et al. 2001).

3.2 Molecular architectures of α-amylases in phylogenetically diverse taxa

Different microorganisms have specific conserved domains and existence of three to four domains was identified as an interesting feature that might be contributing to their adaptation towards differential degradation of starch. These are crucial for functioning of enzymes since they are aligned and distributed at equal intervals within the protein and these regions are responsible for formation of active sites as well as substrate-binding sites (Kuriki and Imanaka 1999). Different types of conserved domains and CBM were present in α-amylase sequences of microbes including bacteria, fungi, algae, cyanobacteria, and higher plants. Moreover, some organisms also showed the presence of multiple copies of CBMs along with unique amylase domains while some consisted of only one type of domain (Supplementary table 4). The molecular architecture of α-amylase sequences present in diverse groups of microbes may help us to understand the evolutionary trends and predict the closely related species.

Among photosynthetic microbes like algae, the PLN02784 domain family was found in eleven species out of which *Auxenochlorella protothecoides* (tr-A0A3M7KWR6, tr-A0A1D2AH16 and tr-A0A087SDV7), *Micractinium conductrix* (tr-A0A2P6VLF6), and *Chlorella variabilis* (tr-E1ZET0) had an additional GlnD family. Additionally, PLN02447 family was found in the amino acid sequences of seven algal taxa (tr-C1FDK3, tr-C1MXZ5, tr-Q6PYZ4, tr-A8HW52, tr-A8IHX1, tr-A0A087STK7, tr-A0A2P6U2U4). All the members of cyanobacteria possess AmyAc domain in their molecular design. Non-photosynthetic microbes such as fungi also consisted of AmyAc domain with some exceptions. For instance; PLN02447 family is present in *Emericella nidulans* (sp-Q9Y8H3), *Saccharomyces cerevisiae* (tr-A6ZQT8 and tr-C7GX32), and *Fusarium oxysporum* (tr-A0A559KXC2). *Aspergillus flavus* consisted of hydrolytic domain 28 (GH28) along with PLN02447 and PRK10118 families. The α-amylase sequences present in *Arabidopsis thaliana* have different domains like PLN02447, PLN02784, PLN00196, and PLN03244. The presence of PLN02447 and PLN02784 families depicts that these may share a common ancestry with algal taxa (Fig. 1). AmyAc superfamily is present in a majority of bacterial species except for *Bacillus* sp. (tr- A0A328L7B1), *Pyrococcus furiosus* (sp-P49067), *Pyrococcus abyssi* (sp-Q9V298), and *Saccharolobus solfataricus* (tr- P95869). Interestingly; in *Bacillus circulans* (sp-A0P8X0) two different types of CBM domains i.e., CBM 25 and CBM 20 were identified. *Bacillus licheniformis* (sp-Q04977) consisted additional α-amylase domain along with AmyAc family, whereas, *Alicyclobacillus* sp. (tr- F2VRZ2) had multiple Eset_CDase domains and Malt amylase C and AmyAc super-families, respectively. The presence of distinct domains in the above bacterial species describes the phylogenetic trend among the organisms as they evolved separately without forming clusters (Fig. 1). The presence of additional Malt amylase C domain was predicted in most of the bacterial species like *Bacillus amyloliquefaciens*, *Bacillus subtilis*, *Geobacillus* sp., and *Bacillus megaterium*. Out of nine α-amylase sequences present in different members of *Streptomyces* sp., seven sequences consisted of CBM 20 along with Malt amylase C and AmyAc superfamily. *Micromonospora haikouensis* (tr-A0A0D0WST3) and *Micromonospora* sp. (tr-A0A3E2YK12) had two copies of CBM 20.

Higher production of α-amylase protein in certain microorganisms may be due to the occurrence of multiple copies of either identical or dissimilar carbohydrate-binding motifs. This may be a cause behind the varying substrate-binding capability of enzymes and probably directed towards adaptive evolution. Literature survey showed that different strains of same species are able to produce variable amounts of α-amylase like in a study, *B. subtilis* RSKK96 strain produced up to 858.6IUmg^{-1} (Akcan et al. 2011) whereas 594 IU g^{-1} of α-amylase was obtained from *B. subtilis* D19 (Almanaa et al. 2020). *B. subtilis* IP 5832 was capable in producing 2.5 IU mL^{-1} of α-amylase (Božić et al. 2011). High α-amylase activity was reported in *B. subtilis* SUNGB2 (22.14 IUmL^{-1}) and *Bacillus licheniformis* HULUB1 (18.15 IUmL^{-1}), respectively (Msarah et al. 2020). Variable production efficiencies among diverse organisms maybe due to distinct amino acids present in the sequences. Mutations in their catalytic sites and carbohydrate-binding modules may alter the starch binding and degradation patterns by different taxa because of their fine-tuned gene expression profiles. Absence of carbohydrate-binding modules among organisms can also be a possible reason for functional inactivity of α-amylase.

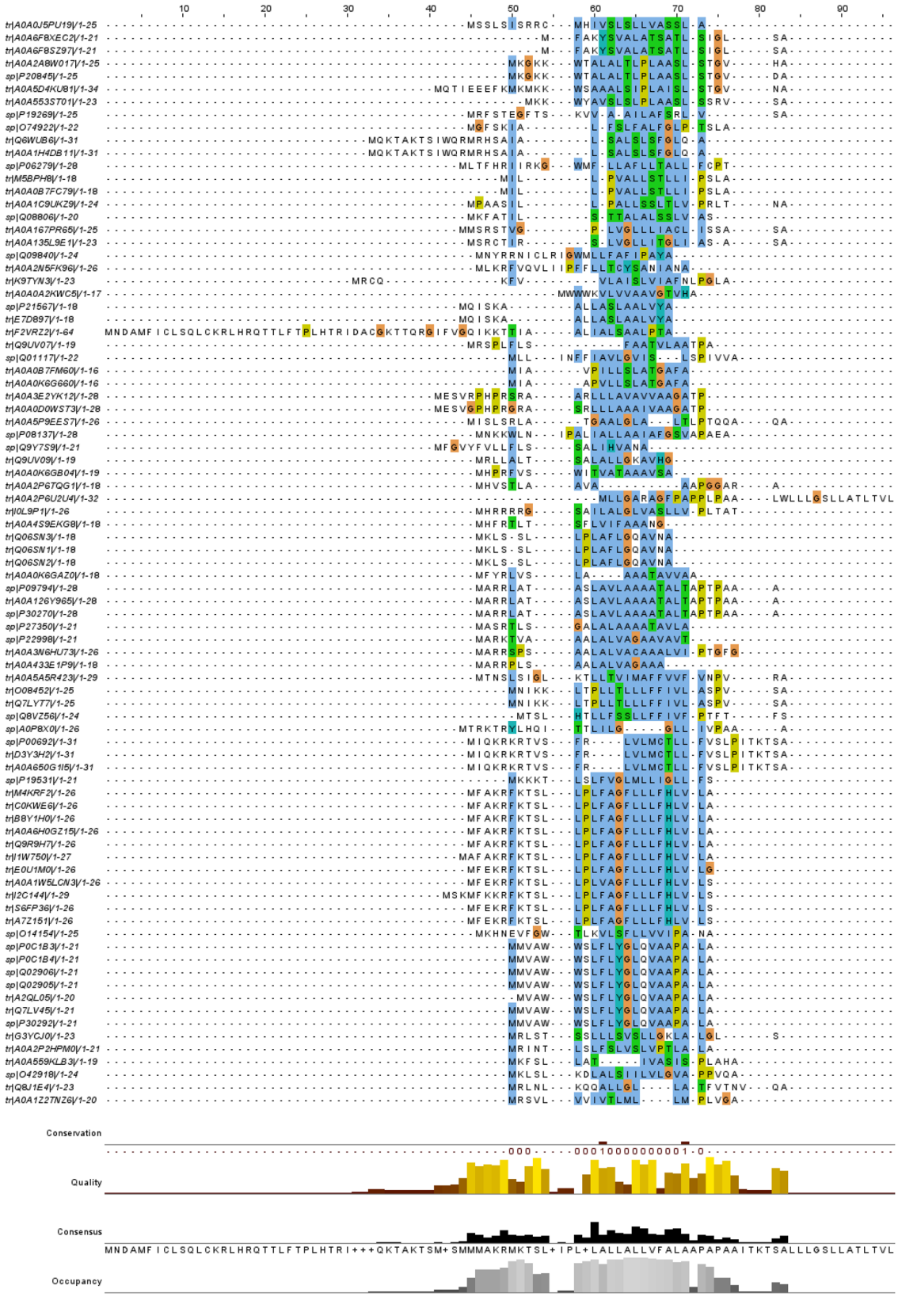

Fig. 1 Multiple sequence alignment of different secretory peptides retrieved from α-amylase amino acid sequences of diverse organisms displaying mutations present in N and C regions

3.3 Identification of secretory peptides

Signal peptides are small-sequences of usually positive-charged amino acids present at N-terminus of secreted proteins which mediates intracellular to extracellular transport of protein from the cell to outside where they perform their functions. Several variations among amino termini of α-amylase sequences and different types of signal peptides were identified from prokaryotic and eukaryotic microorganisms. Different microorganisms are known to possess different amylolytic potential which may significantly depend upon the amino acids present in signal peptide sequence of α-amylase. Positively charged amino acid residues present at N-terminus are known to interact with translocation mechanisms (Freudl 2018; Wu et al. 2020). Results from MSA indicate the presence of variations among microorganisms at N and C regions of secretory peptides (Fig. 1). Thus, the secretory capacity of amino acid residues may vary depending on the mutations present. Out of 197 sequences, we found 85 signal peptides from diverse α-amylase sequences involved in this study (Supplementary Table 5).

Among α-amylase sequences present in green algae, we observed signal peptides in two sequences of *Chlorella sorokiniana* of dissimilar lengths. In amino acid sequences of cyanobacteria; two sequences i.e., *Chroococcidiopsis thermalis* and *Microcystis aeruginosa* showed the presence of signal peptides. Majority of algae and cyanobacteria included in this study did not possess any secretory peptides. Additionally, CBM was lacking in algae and cyanobacteria that alternatively may be a cause to produce inactive α-amylase. Out of 46 α-amylase sequences of fungi, signal peptides could be predicted in 37 sequences only. *Aspergillus oryzae* (sp-P0C1B3 and sp-P0C1B4), *Aspergillus awamori* (sp-Q02906 and sp-Q02905), *Aspergillus flavus* (tr-Q7LV45), and (*Aspergillus usamii* sp-P30292) have same secretory peptides. *Aspergillus niger* (tr-A2QL05) also have a similar peptide sequence like the above microbes but lacked a methionine residue when compared to the above peptides of fungi. In contrast to other *Aspergillus* sp.; *Aspergillus niger* (tr-G3YCJ0), *Aspergillus fumigatus* Z5 (tr-A0A0J5PU19), and *Aspergillus flavus* (tr-A0A2P2HPM0) had different signal peptide residues. *Penicillium expansum, Penicillium chrysogenum*, and *Penicillium patulum* have different signal peptides in α-amylase sequences. In case of *Rhizoctonia solani*, we observed four distinct types of signal peptides and therefore, inferred that alterations in amino acid residues may cause variations in enzyme secreting capabilities of different organisms. *Ophiostoma floccosum and Saccharomycopsis fibuligera* had similar peptides whereas, in case of *Schizosaccharomyces pombe*, *Schwanniomyces occidentalis*, *Lipomyces kononenkoae* presence of different types of signal peptides with varying lengths could be predicted. Among four sequences of *Arabidopsis thaliana*, only one sequence showed the existence of secretory peptides. Out of 89 bacterial α-amylase sequences, we found that 43 sequences consisted of secretory peptides. *Bacillus* sp. and *Bacillus amyloliquefaciens* have similar types of peptides consisting of thirty-one amino acid residues. In case of *Bacillus amyloliquefaciens*, we found two distinct peptides of dissimilar lengths. Among *Bacillus velezensis* (tr-S6FP36 and tr-A7Z151) and *B. amyloliquefaciens* (tr-A0A1W5LCN3), we observed that signal peptide sequences were similar, however, in tr-A0A1W5LCN3; valine was present instead of phenylalanine when compared with above two sequences. *Bacillus subtilis* have similar patterns of signal peptide sequences but in tr-I1W750 we found additional alanine residue whereas in tr-E0U1M0; glutamatic acid and glycine were identified instead of two alanine residues. Two different signal peptide sequences were found in *Bacillus megaterium*. In halophilic bacterium, *Halomonas meridiana*, similar types of signal peptide sequences were predicted. However, in case of *Geobacillus thermoleovorans* and *Thermococcus sp.* two distinct peptides were found. Three types of secretory peptides were identified in *Bacillus circulans* with variable lengths. *Alicyclobacillus* sp. had sixty-four residues secretory-peptide, which was longest in length among all the peptide sequences identified. In actinobacteria, *Streptomyces* sp., nine secretory peptides were found out of which *Streptomyces limosus* (sp-P09794) and *Streptomyces albidoflavus* (tr-A0A126Y965) had similar peptide sequence whereas distinct signal peptides were present in other species. Functionally inactive α-amylase present in algae, cyanobacteria, certain members of fungi, and bacteria may be due to the absence of signal peptides or variations in its secretory-peptide sequences.

Therefore, inferences made from the study showed that different groups of microbes possessed different kinds of signal peptides. Interestingly, even in same group or different organisms of same genus variations were found. As identified from the published database discussed above for *Bacillus* strains having different enzymatic activities, however, they belonged to same genus but they may have different secretory peptides. Similarly, in our findings even in same genus variations were found among secretory peptides which alternatively predict their varying functioning.

3.4 Evolutionary tree

We tried to understand the molecular progression among organisms with distinct α-amylase sequences through phylogenetic analysis. The amino acid residues present in the sequence were used to interpret the evolutionary relatedness among a wide range of algae, cyanobacteria, bacteria, fungi, and plants. Gaps were also included in multiple sequence alignments because these gaps revealed about the mutations that happened during evolution. These mutations are referred

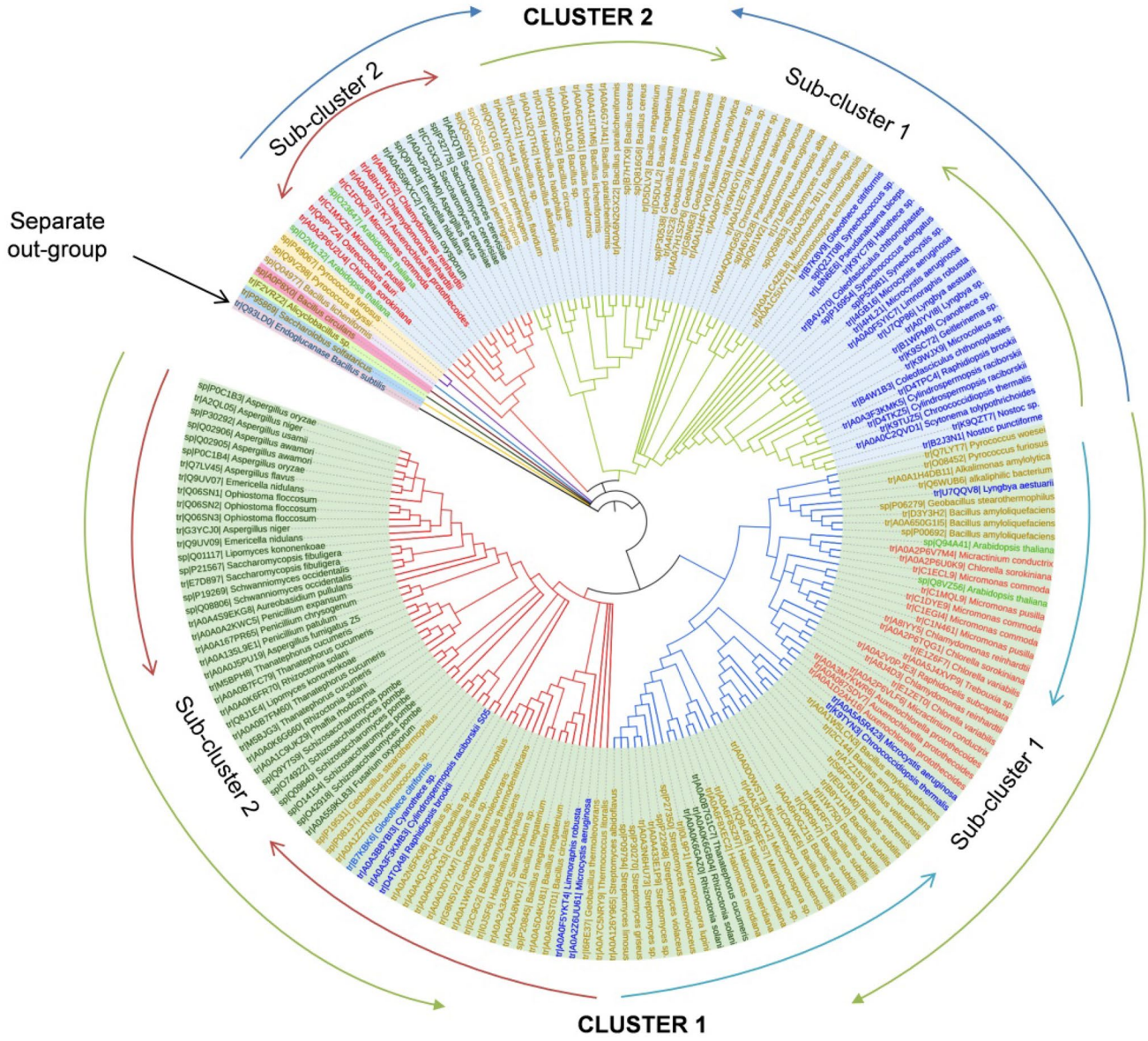

Fig. 2 Phylogenetic tree of 197 α-amylase sequences present in diverse group of microorganisms. Construction of tree was done using bootstrap method with 500 replications. Endoglucanase sequence of *Bacillus subtilis* (tr-Q93LD0) was used as an outgroup sequence

to as "*InDel*" mutations because of insertion and deletion of amino acid residues (Simmons and Ochoterena 2000). The comprehensive distribution of all 197 α-amylase sequences is represented in the phylogenetic tree (Fig. 2). In addition, to study the similarity in various amino acid sequences, an outgroup sequence of endoglucanase from *Bacillus subtilis* (tr-Q93LD0) was included which separated all the organisms into two major clusters. The evolutionary tree is distributed in two major clusters splitting into smaller sub-clusters of algae, bacteria, cyanobacteria, fungi, and *Arabidopsis thaliana*. The first major cluster contained fewer cyanobacteria taxa compared to the second. The second cluster mainly comprises of bacterial and cyanobacterial species. Compared to the first cluster, less number of fungal and algal species were observed in second cluster. First major cluster have two sub-clusters which was further divided into sub various smaller clusters. *Lyngbya aestuarii* (tr-U7QQV8) shared a common ancestry with *Geobacillus stearothermophilus* (sp-P06279) and *Bacillus amyloliquefaciens* (tr-D3Y3H2, tr- A0A650G1I5 and sp- P00692). *Arabidopsis thaliana* (sp-Q94A41) shared common ancestral history with tr-A0A2P6V7M4, tr-A0A2P6U0K9 & tr-C1ECL9 (Fig. 2).

Furthermore, in second cluster, two sub-clusters were formed which were further divided into smaller sub-clusters representing evolution in various groups of microorganisms. In sub-cluster two, we observed that *Arabidopsis thaliana*

(sp-O23647) shared a similar evolutionary pattern with tr-A8HW52, tr-A8IHX1, tr-A0A087STK7 and tr-C1FDK3. It was found that members of GH57 family, *Pyrococcus furiosus* (sp-P49067) and *Pyrococcus abyssi* (sp-Q9V298) were evolved as separate clade apart from major cluster 1 and 2. *Geobacillus stearothermophilus* (tr-A0A0K2HA33), another member of GH57 family was grouped with α-amylase proteins of GH13 family, suggesting common ancestry. The α-amylase protein of *Bacillus circulans* (sp-A0P8X0) belonging to GH119 family, evolved as separate lineage in the evolutionary tree (Fig. 2), indicating that instead of having a common family (GH119), it must have some unique sequence pattern than others because of which it evolved separately.

To study the evolution of signal peptides, all the sequences were annotated for the presence or absence of secretory sequences which were later used for phylogenetic analysis. The α-amylases have distinct types of secretory peptides present which may influence the secretion levels of the enzymes. We retrieved signal peptides from α-amylase sequences of diverse organisms and aligned them by multiple sequence alignment. Further, an evolutionary tree was generated to understand the evolutionary pattern (Fig. 3). Most of the organisms have varying secretion capacity when compared to others; therefore, tree construction may help in understanding their ancestral history. Organisms that did not possess secretory peptides were excluded from the study. From the evolutionary tree, we observed that various clusters were formed whereas twenty-two signal peptide sequences of various microorganisms evolved as separate lineage. Presence of variations in the amino acid residues of signal peptides could be a possible reason for evolution towards differential secretion patterns. The first cluster consisted of eleven members of *Bacillus* sp. and one *Geobacillus stearothermophilus* which depicts that they share a common ancestry. The second cluster included six members of *Aspergillus* sequences. Three sequences of *Streptomyces* species and two of fungi (*Schizosaccharomyces pombe* and *Lipomyces kononenkoae*) were assorted in third cluster. The fourth,

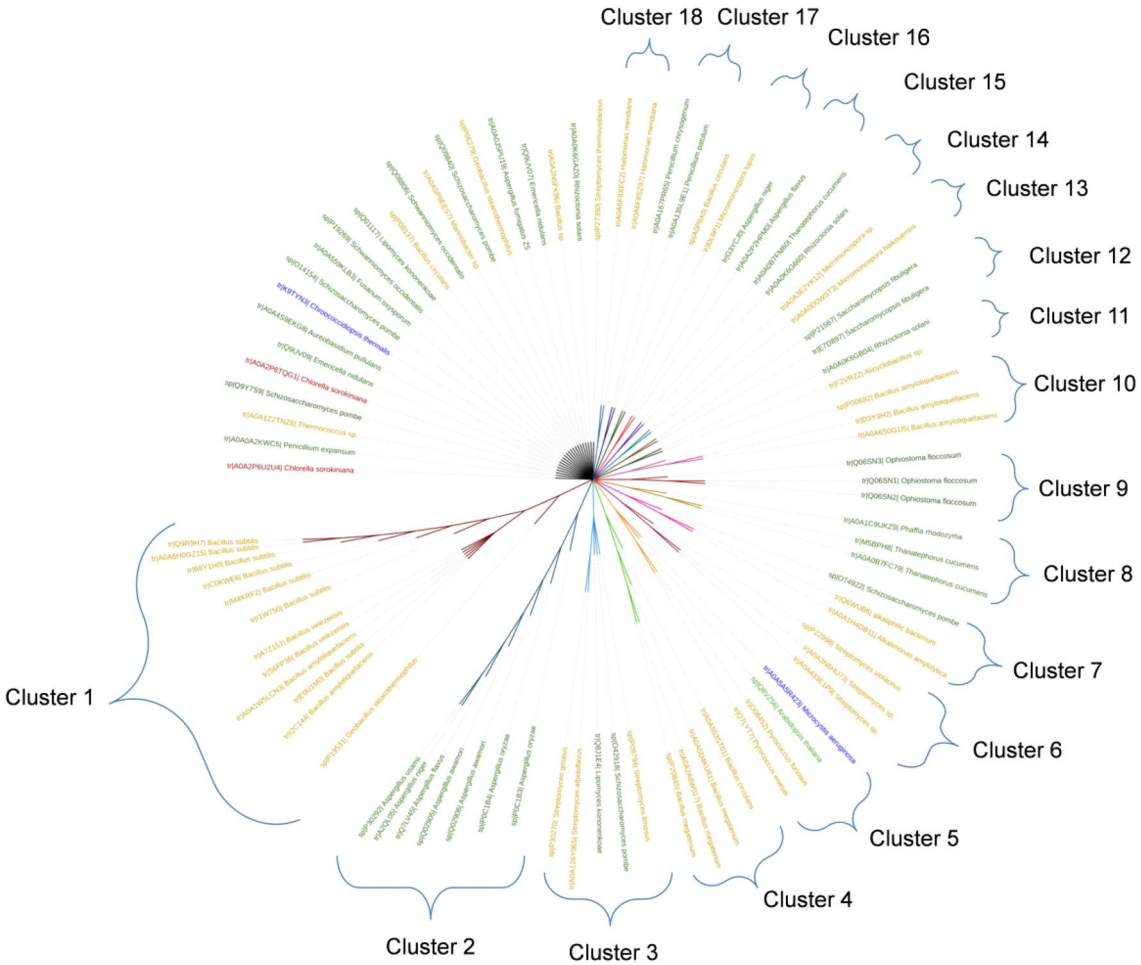

Fig. 3 Unrooted phylogenetic tree of signal peptides present in α-amylase sequences of diverse microbes

sixth, tenth, thirteen, sixteen, and eighteen cluster comprised bacterial species only. The fifth cluster consisted *of Pyrococcus woesei, Pyrococcus furiosus, Microcystis aeruginosa*, and *Arabidopsis thaliana* whereas in seventh cluster; *Schizosaccharomyces pombe* was present with *Alkalimonas amylolytica* and *alkaliphilic bacterium*. The small clusters (eighth, ninth, twelve, fourteen, fifteen and seventeen) consisted of fungal organisms and the eleventh cluster had *Alicyclobacillus* sp. with *Rhizoctonia solani*.

Based on the observations, it was found that secretory peptides have varying evolutionary patterns. Usually, they had consistencies among a group of microbes, however, in some cases, signal peptides of different groups of microbes evolved simultaneously as a common clade indicating that they gained the unique mutations which helped them to assort with diverse groups of taxa. For instance, in cluster 5 (Fig. 3) signal peptides of phylogenetically diverse taxa of archaeabacteria; *Pyrococcus furiosus*, plant; *Arabidopsis thaliana* and cyanobacteria; *Microcystis aeruginosa* were evolved simultaneously. In case of different signal peptide sequences of various taxa of genus *Bacillus,* an interesting evolutionary trend was identified. Among all the signal peptides evaluated, secretory signal sequences of *Bacillus* showed vast diversity and were assorted in different clusters. Similarly, signal peptides of yeast sequences also showed diverse phylogenetic trends.

4 Conclusions

The *in-silico* approaches and conservation pattern among α-amylase sequences obtained from proteomic database highlight the evolutionary relatedness among closely related groups of microorganisms including photosynthetic and non-photosynthetic members. Based on their molecular architectures and conserved domains, phylogenetically diverse organisms evolved as distinct clusters. Higher plants, algae, and cyanobacteria lacked signal peptides and carbohydrate-binding domains. This may affect the amylolytic activity of α-amylase enzyme; therefore, it may contribute to non-functionality of the enzyme. In case of fungi and bacteria, most of the species consisted of multiple carbohydrate-binding domains along with phylogenetically diverse secretory peptides which make them potential candidates for differential α-amylase production. Increased demands for enzyme production from microbial sources are beneficial to the industries due to decreased processing time, lower costs and environment friendly attributes. This study helped to understand the divergence and functional dissimilarities of α-amylases among various taxonomic groups. Additionally, this study helped to know more about the major taxonomic groups having functional α-amylases. These computational tools may help researchers to predict the concept behind varying enzyme evolution among different organisms. This knowledge can alternatively be utilized to re-design the suitable genetic constructs with desired signal peptides, multiple copies of carbohydrate-binding domains and unique hydrolytic domains for improved enzyme production.

5 Future perspectives

Novel approaches including protein engineering, cloning, and expression studies in combination with phylogenetic analysis from variable plant and microbial sources may help in understanding the significant features related to the evolution of α-amylase enzyme. Phylogenetic studies help in designing the enzyme architecture by identifying distinct CBMs and signal peptides present in variable organisms. The knowledge about domain distribution may help in understanding the catalytic activity and overall secretion capacity of microorganisms. Bioinformatics tools can be used to construct a chimeric enzyme containing multiple domains for increased substrate degradation of polysaccharides. Alterations in the domain arrangement can create chimeric enzymes with enhanced enzymatic activity and these studies offer new strategies for developing multiple domain enzymes. Additionally, knowledge about the organization of hydrolytic domains, CBMs, and secretory peptides is a significant approach towards enzyme engineering. These evolutionary studies may further improve the characteristics of enzymes and enhance the production yields of enzymes required in industrial applications. The modern technologies with computational studies may also help the researchers to explore potential α-amylases from different sources and meet the demands of industrial sector.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12210-022-01068-7.

Acknowledgements The authors greatly acknowledge R and D division of Tata Steel Ltd, Jamshedpur for providing infrastructural and all the required facilities.

Author contributions SK, VKP and RK designed the experiments, SK, RK drafted the manuscript; SK, VKP, AD, and RK revised and approved the manuscript.

Funding Tata Steel.

Declarations

Conflict of interest The authors declare that they have no competing interests.

References

Akcan N, Fikret U, Güven A (2011) Alpha-amylase production by *Bacillus subtilis* RSKK96 in submerged cultivation. Kafkas Üniversitesi, Veteriner Fakültesi Dergisi 17:17–22

Almanaa TN, Vijayaraghavan P, Alharbi NS, Kadaikunnan S, Khaled JM, Alyahya SA (2020) Solid state fermentation of amylase production from *Bacillus subtilis* D19 using agro-residues. J King Saud Univ Sci 32(2):1555–1561. https://doi.org/10.1016/j.jksus.2019.12.011

Balakrishnan M, Jeevarathinam G, Kumar SKS, Muniraj I, Uthandi S (2021) Optimization and scale-up of α-amylase production by *Aspergillus oryzae* using solid-state fermentation of edible oil cakes. BMC Biotechnol 21(1):1–11

Bayer EA, Morag E, Wilchek M, Lamed R, Yaron S, Shoham Y (1995) Cellulosome domains for novel biotechnological application. In: Progress in biotechnology, vol 10. Elsevier, Amsterdam, Netherland, pp 251–259

Bernardes A et al (2019) Carbohydrate binding modules enhance cellulose enzymatic hydrolysis by increasing access of cellulases to the substrate. Carbohydr Polym 211:57–68

Bhatia S (2018) Introduction to pharmaceutical biotechnology, Volume 2; Enzymes, proteins and bioinformatics. ISBN: 978-0-7503-1302-5

Boraston AB, Nurizzo D, Notenboom V, Ducros V, Rose DR, Kilburn DG, Davies GJ (2002) Differential oligosaccharide recognition by evolutionarily-related β-1, 4 and β-1, 3 glucan-binding modules. J Mol Biol 319:1143–1156

Božić N, Ruiz J, López-Santín J, Vujčić Z (2011) Optimization of the growth and α-amylase production of *Bacillus subtilis* IP 5832 in shake flask and laboratory fermenter batch cultures. J Serb Chem Soc 76(7):965–972

Choudhuri S (2014) Bioinformatics for beginners: genes, genomes, molecular evolution, databases and analytical tools. Elsevier

Choudhury AKR (2020) Introduction to enzymes. In: Sustainable technologies for fashion and textiles. Elsevier, Amsterdam, Netherland, pp 75–90 https://doi.org/10.1016/B978-0-08-102041-8.00010-X

Cockburn DW, Suh C, Medina KP, Duvall RM, Wawrzak Z, Henrissat B, Koropatkin NM (2018) Novel carbohydrate binding modules in the surface anchored α-amylase of *Eubacterium rectale* provide a molecular rationale for the range of starches used by this organism in the human gut. Mol Microb 107:249–264

Coutinho P. (1999) Carbohydrate-active enzymes: an integrated database approach. Recent advances in carbohydrate bioengineering. Proceedings of the 3rd Carbohydrate Bioengineering Meeting, University of Newcastle upon Tyne, UK, 11–14 April 1999: 3–12 https://eurekamag.com/research/003/375/003375840.php

El-Enshasy HA, Abdel Fattah YR, Othman NZ (2013) Amylases: characteristics, sources, production, and applications bioprocessing technologies in biorefinery for sustainable production of fuels, chemicals, and polymers. Wiley, USA https://doi.org/10.1002/9781118642047.ch7

Farber GK, Petsko GA (1990) The evolution of α/β barrel enzymes. Trends Biochem Sci 15(6):228–234. https://doi.org/10.1016/0968-0004(90)90035-A

Ficko-Blean E, Stuart CP, Boraston AB (2011) Structural analysis of CPF_2247, a novel α-amylase from *Clostridium perfringens.* proteins: structure. Func, Bioinform 79:2771–2777. https://doi.org/10.1002/prot.26325

Freudl R (2018) Signal peptides for recombinant protein secretion in bacterial expression systems. Microb Cell Fact 17(1):1–10. https://doi.org/10.1186/s12934-018-0901-3

Fukusumi S, Kamizono A, Horinouchi S, Beppu T (1988) Cloning and nucleotide sequence of a heat-stable amylase gene from an anaerobic thermophile *Dictyoglomus thermophilum*. Europ J Biochem 174(1):15–21

Gangadharan D, Jose A, Nampoothiri KM (2020) Recapitulation of stability diversity of microbial α-amylases. Amylase 4:11–23. https://doi.org/10.1515/amylase-2020-0002

Geer LY, Domrachev M, Lipman DJ, Bryant SH (2002) CDART: protein homology by domain architecture. Genome Res 12(10):1619–1623

Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 313:903–919

Henrissat B, Bairoch A (1993) New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem J 293(3):781–788

Hiller K, Grote A, Scheer M, Münch R, Jahn D (2004) PrediSi: prediction of signal peptides and their cleavage positions. Nucleic Acids Res 32:W375–W379

Hmidet N, Ali N, Haddar A, Kanoun S, Sellami-Kamoun A (2009) Alkaline proteases and thermostable α-amylase co-produced by *Bacillus licheniformis* NH1: Characterization and potential application as detergent additive. Biochem Eng J 47:71–79

Janeček Š, Gabriško M (2016) Remarkable evolutionary relatedness among the enzymes and proteins from the α-amylase family. Cell Mol Life Sci 73:2707–2725

Janeček Š, Svensson B (2022) How many α-amylase GH families are there in the CAZy database? Amylase 6:1–10. https://doi.org/10.1515/amylase-2022-0001

Janeček Š, Svensson B, MacGregor E (2014) α-Amylase: an enzyme specificity found in various families of glycoside hydrolases. Cell Mol Life Sci 71:1149–1170

Janeček Š, Mareček F, MacGregor EA, Svensson B (2019) Starch-binding domains as CBM families–history, occurrence, structure, function and evolution. Biotechnol Adv 37:107451

Kumar D, Singh V (2016) Dry-grind processing using amylase corn and superior yeast to reduce the exogenous enzyme requirements in bioethanol production. Biotechnol Biofuels 9(1):1–12

Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35(6):1547–1549

Kuriki T, Imanaka T (1999) The concept of the α-amylase family: structural similarity and common catalytic mechanism. J Biosci Bioeng 87(5):557–565. https://doi.org/10.1016/S1389-1723(99)80114-5

Laderman K, Asada K, Uemori T, Mukai H, Taguchi Y, Kato I, Anfinsen C (1993) Alpha-amylase from the hyperthermophilic archaebacterium *Pyrococcus furiosus*. cloning and sequencing of the gene and expression in *Escherichia coli*. J Biol Chem 268:24402–24407

Letunic I, Bork P (2021) Interactive tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49(W1):W293–W296. https://doi.org/10.1093/nar/gkab301

Li F, Zhu X, Li Y, Cao H, Zhang Y (2011) Functional characterization of a special thermophilic multifunctional amylase OPMA-N and its N-terminal domain. Acta Biochim Biophys Sin 43:324–334

Linden A, Mayans O, Meyer-Klaucke W, Antranikian G, Wilmanns M (2003) Differential regulation of a hyperthermophilic alpha-amylase with a novel (Ca, Zn) two-metal center by zinc. J Biol Chem 278(11):9875–9884. https://doi.org/10.1074/jbc.M211339200

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

Lu S et al (2020) CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res 48(D1):D265–D268

Madeira F et al (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res 47(W1):W636–W641. https://doi.org/10.1093/nar/gkz268

Marchler-Bauer A, Bryant SH (2004) CD-Search: protein domain annotations on the fly. Nucleic Acids Res 32:W327–W331

Marchler-Bauer A et al (2010) CDD: a Conserved domain database for the functional annotation of proteins. Nucleic Acids Res 39:D225–D229

Martin M, Okpo E, Andy I (2019) Microbial amylases: a review. World News Nat Sci 22:174–179

Mehta D, Satyanarayana T (2016) Bacterial and archaeal α-amylases: diversity and amelioration of the desirable characteristics for industrial applications. Front Microb 7:1129

Msarah MJ, Ibrahim I, Hamid AA, Aqma WS (2020) Optimisation and production of alpha amylase from thermophilic *Bacillus spp*. and its application in food waste biodegradation. Heliyon 6(6):e04183

Mukherjee AK, Borah M, Rai SK (2009) To study the influence of different components of fermentable substrates on induction of extracellular α-amylase synthesis by *Bacillus subtilis* DM-03 in solid-state fermentation and exploration of feasibility for inclusion of α-amylase in laundry detergent formulations. Biochem Eng J 43(2):149–156

Muralikrishna G, Nirmala M (2005) Cereal α-amylases—an overview. Carbohydr Polym 60:163–173

Nielsen JE, Borchert TV (2000) Protein engineering of bacterial α-amylases. Biochimica et Biophysica Acta (BBA) Protein Struc Mol Enzymol 1543:253–274

Nielsen MM et al (2009) Two secondary carbohydrate binding sites on the surface of barley α-amylase 1 have distinct functions and display synergy in hydrolysis of starch granules. Biochemistry 48:7686–7697

Paul D (2016) Microorganisms and α-amylase: a concise review. Inv J Sci 4:1–5

Podani J, Morrison DA (2017) Categorizing ideas about systematics: alternative trees of trees, and related representations. Rend Fis Acc Lincei 28:191–202

Robinson PK (2015) Enzymes: principles and biotechnological applications. Essays Biochem 59:1–41

Sidar A, Albuquerque ED, Voshol GP, Ram AF, Vijgenboom E, Punt PJ (2020) Carbohydrate binding modules: diversity of domain architecture in amylases and cellulases from filamentous microorganisms. Front Bioeng Biotechnol. https://doi.org/10.3389/fbioe.2020.00871

Sievers F et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Mol Syst Biol 7(1):539

Simmons MP, Ochoterena H (2000) Gaps as characters in sequence-based phylogenetic analyses. Syst Biol 49(2):369–381. https://doi.org/10.1093/sysbio/49.2.369

de Souza Vandenberghe LP et al. (2020) Classification of enzymes and catalytic properties. In: Biomass, biofuels, biochemicals. Elsevier, Amsterdam, Netherland, pp 11–30 https://doi.org/10.1016/B978-0-12-819820-9.00002-8

Stam MR, Danchin EG, Rancurel C, Coutinho PM, Henrissat B (2006) Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of α-amylase-related proteins. Protein Eng Des Sel 19:555–562

Tiwari S et al (2015) Amylases: an overview with special reference to alpha amylase. J Global Biosci 4:1886–1901

Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 307(4):1113–1143

UniProt: the universal protein knowledgebase in 2021 (2021) Nucleic Acids Res. 49(D1):D480–D489 https://doi.org/10.1093/nar/gkaa1100

Watanabe H, Nishimoto T, Kubota M, Chaen H, Fukuda S (2006) Cloning, sequencing, and expression of the genes encoding an isocyclomaltooligosaccharide glucanotransferase and an α-amylase from a *Bacillus circulans* strain. Biosci Biotechnol Biochem 70:2690–2702

Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189–1191

Wu Z et al (2020) Signal peptides generated by attention-based neural networks ACS Synth. Biol 9:2154–2161. https://doi.org/10.1021/acssynbio.0c00219