The starch-binding domain family CBM41 - an in silico analysis of evolutionary relationships

Carbohydrate-binding module family 41

Janeek, Štefan; Majzlová, Katarína; Svensson, Birte; MacGregor, E. Ann

Published in:
Proteins: Structure, Function, and Bioinformatics

Link to article, DOI:
10.1002/prot.25309

Publication date:
2017

Document Version
Peer reviewed version

Citation (APA):
Janeek, Š., Majzlová, K., Svensson, B., & MacGregor, E. A. (2017). The starch-binding domain family CBM41 - an in silico analysis of evolutionary relationships: Carbohydrate-binding module family 41. Proteins: Structure, Function, and Bioinformatics, 85(8), 1480-1492. DOI: 10.1002/prot.25309

DOI 10.1002/prot.25309

The starch-binding domain family CBM41 – an *in silico* analysis of evolutionary relationships

Štefan Janeček,[1,2]* Katarína Majzlová,[1] Birte Svensson,[3] and E. Ann MacGregor[4]

[1] Institute of Molecular Biology, Slovak Academy of Sciences, Bratislava, Slovakia
[2] Department of Biology, Faculty of Natural Sciences, University of SS. Cyril and Methodius, Trnava, Slovakia
[3] Department of Biotechnology and Biomedicine, Technical University of Denmark, Kgs. Lyngby, Denmark
[4] 2 Nicklaus Green, Livingston, West Lothian, United Kingdom

ABSTRACT

Within the CAZy database, there are 81 carbohydrate-binding module (CBM) families. A CBM represents a non-catalytic domain in a modular arrangement of glycoside hydrolases (GHs). The present *in silico* study has been focused on starch-binding domains from the family CBM41 that are usually part of pullulanases from the α-amylase family GH13. Currently there are more than 1,600 sequences classified in the family CBM41, almost exclusively from *Bacteria*, and so a study was undertaken in an effort to divide the members into relevant groups (subfamilies) and also to contribute to the evolutionary picture of family CBM41. The CBM41 members adopt a β-sandwich fold (~100 residues) with one carbohydrate-binding site formed by the side-chains of three aromatic residues that interact with carbohydrate. The family CBM41 can be divided into two basic subdivisions, distinguished from each other by a characteristic sequence pattern or motif of the three essential aromatics as follows: (i) "W-W-~10aa-W" (the so-called *Streptococcus*/*Klebsiella*-type); and (ii) "W-W-~30aa-W" (*Thermotoga*-type). Based on our bioinformatics analysis it is clear that the first and second positions of the motif can be occupied by aromatic residues (Phe, Tyr, His) other than tryptophan, resulting in the existence of six different carbohydrate-binding CBM41 groups, that reflect mostly differences in taxonomy, but which should retain the ability to bind an α-glucan. In addition, three more groups have been proposed that, although lacking the crucial aromatic motif, could possibly employ other residues from remaining parts of their sequence for binding carbohydrate.

Key words: starch-binding domain; carbohydrate-binding module family CBM41; family GH13 pullulanases; protein sequence-structural comparison; evolutionary relatedness.

Abbreviations: CAZy, Carbohydrate-Active enZymes; CBM, carbohydrate-binding module; GBD, glycogen-binding domain; GH, glycoside hydrolase; PDB, Protein Data Bank; RMSD, root-mean square deviation; SBD, starch-binding domain.

Running title: Carbohydrate-binding module family 41

* Correspondence to: Štefan Janeček, Laboratory of Protein Evolution, Institute of Molecular Biology, Slovak Academy of Sciences, Dúbravská cesta 21, SK-84551 Bratislava, Slovakia.
E-mail: Stefan.Janecek@savba.sk

Received: Dec 02, 2016; Revised: Apr 05, 2017; Accepted: Apr 17, 2017

INTRODUCTION

The carbohydrate-binding module (CBM) family 41 is one of the 81 CBM families classified currently within the Carbohydrate-Active enZymes (CAZy; http://www.cazy.org/) database.[1] This module folds as a single domain and, for the characterized enzymes, has been found to exist together with the catalytic domains of pullulanases and glycogen-degrading enzymes[2,3] of glycoside hydrolase (GH) family GH13, the main α-amylase family.[4,5] This view is supported by the first crystal structure of a CBM41 module determined in complex with maltooligosaccharides in the family GH13 pullulanase from *Thermotoga maritima*.[6] The CBM41 is therefore considered as one of the 13 families of starch binding domains (SBDs), i.e., CBM20, 21, 25, 26, 34, 45, 48, 53, 58, 68, 69 and 74 in addition to CBM41.[1] In some cases a domain, classified as an SBD, may actually represent a glycogen-binding domain (GBD), e.g., in laforin, genethonin-1 or the β-subunit of AMP-activated protein kinase.[7]

In general, an SBD is a protein module consisting of approximately 100 amino acid residues and containing aromatic residues at certain positions (mostly tryptophans and tyrosines, but also phenylalanines or histidines) that are essential for the SBD to bind carbohydrate,[7-14] i.e., raw starch and/or related α-glucans, and thus enhance the degradative action of the enzyme to which the SBD is attached.[15,16] It is of note that raw-starch degrading amylases without any distinct SBD exist and are also well known.[17] The individual SBDs (and GBDs) classified as the CAZy CBM families mentioned above can be distinguished from each other by the conserved positions of these aromatic residues that constitute either one or two starch-binding sites. In other words, although the overall tertiary structure of all SBDs (GBDs included) is a similar β-sandwich[13,18,19] and the mode of saccharide binding may be conserved,[7,20-22] the individual SBD families exhibit characteristic aromatic residues involved in binding, which are fully alignable only within a given family.[7-8,12,23]

With regard to the location of an SBD in the domain organization of a polypeptide chain, it may either precede[6,24-32] or succeed the catalytic domain.[33-40] It moreover is frequently found accompanied by so-called fibronectin type III domains occurring in a wide variety of carbohydrate-active enzymes.[41] Remarkably, in a special case, the SBD is inserted within the domain B, which itself represents a long loop in the catalytic $(\beta/\alpha)_8$-barrel domain A, characteristic of all members of the α-amylase family GH13.[42] The N-terminal position is thus typical for families CBM21, 34, 41, 45, 48, 53 and 68, while families CBM20, 25, 26, 69 and 74 occur in a C-terminal position, and the special case of the insertion within domain B involves the family CBM58.[6-12,43-50]

SBDs were originally recognized as a module found typically in microbial amylolytic and related enzymes. It is worth mentioning that various SBDs and GBDs have also been identified in other plant and animal enzymes and proteins, such as plant starch synthase III,[28,51-53] glucan-water dikinases,[26,29,31] starch-excess protein-4[54,55], animal laforin,[56-59] genethonin-1,[60-62] as well as the β-subunit of AMP-activated protein kinase[63,64] and its homologues in plants,[65,66] and even in lytic polysaccharide monooxygenases from fungi.[67] Nevertheless, as for the microbial SBDs, generally the non-microbial SBDs or GBDs are always involved in metabolism of starch and/or glycogen, or other α-glucans.[7]

The family CBM41 was established based on the study by Lammerts van Bueren et al.,[4] who demonstrated binding of α-glucans containing 1,4 glycosidic bonds only (preferred) and with 1,4 and 1,6-bonds (tolerated) by the recombinantly produced SBD of the family GH13 pullulanase from *Thermotoga maritima*. According to the division of the α-amylase family GH13 into subfamilies,[68] this pullulanase[69] belongs to subfamily GH13_14. All the CBM41 family members are likely to adopt a distorted β-barrel fold (β-sandwich) with one carbohydrate-binding site,[6] formed, in *Thermotoga maritima* pullulanase, by the side-chains of two tryptophans W27 and W29 responsible for hydrogen bonding and hydrophobic stacking interaction, respectively, with glucose moieties, and completed by a third tryptophan W73 also involved in a stacking interaction and contributing to binding of longer maltooligosaccharides.

As mentioned above, most CBMs classified as SBDs share a β-sandwich fold[10-14,18,19] using usually aromatic residues at binding sites. It has already been pointed out[6,25] that the overall fold, with limited sequence identity and general architecture of the binding sites, is similar in CBM41, CBM20, CBM25 and CBM26. For the CBM20, it was convincingly shown[7] that it exhibits the best similarity, including both sequence and structural correspondences of binding residues, to CBM48; this pronounced homology has recently been extended also to family CBM69.[40] The two remaining families CBM25 and CBM26 have also been recognized as forming a closely related pair of SBDs.[8,37] The fact that CBM41 is preferably found in pullulanases could be explained by an ability of CBM41 to accommodate oligo- or poly-saccharides with an α-1,6-linked glucose, which CBM20, CBM25 and CBM26 very probably lack.[6] There is no indication, however, that there is a higher affinity for a branched substrate.

Currently, there are more than 1,600 sequences classified in the family CBM41 in the CAZy database.[1] These are almost all from *Bacteria*, with a few members from *Eucarya* – currently only from green and red algae. As mentioned above, with regard to the catalytic domain, family CBM41 modules are mainly contained in pullulanases, glycogen-degrading enzymes and amylopullulanases, all classified in the α-amylase family GH13.[4,5] However, in addition, many putative "amylases", related enzymes and other hypothetical proteins, are predicted to have a CBM41.[1] In addition to the CBM41 from the *Thermotoga maritima* pullulanase,[6] tertiary structures are known for the CBM41 of pullulanases from *Klebsiella pneumoniae*[25] and *Klebsiella oxytoca*,[70] both classified in the subfamily GH13_13. Moreover, in the solved structure of *Streptococcus pneumoniae* glycogen degrading enzyme SpuA from the subfamily GH13_12, two CBM41s are arranged in tandem, importantly in complex with a maltooligosaccharide.[71] Further, the structure of its counterpart, the surface-anchored pullulanase from *Streptococcus pyogenes* PulA, has been determined and shown to possess two CBM41s in tandem.[3]

One goal of the present bioinformatics study was to attempt to divide the sequences of family CBM41 into groups that could potentially define subfamilies. The individual groups exhibit their particular sequence fingerprints, i.e. a pattern (or motif) of aromatic residues involved in carbohydrate binding, which are arranged at specific positions in the amino acid sequence and for which characteristic replacement of two of the three essential tryptophans by other aromatic amino acids are seen. Important further aims were to elucidate the evolutionary relationships among the members of the family CBM41 and to try to relate the observed different patterns, i.e. motifs of binding residues, to substrate specificities represented by subfamilies of GH13, as well as to taxonomy.

MATERIALS AND METHODS

Sequence collection

All sequences containing a family CBM41 (for details, see Table SI) were collected based on information in the CAZy database (http://www.cazy.org/),[1] and completed by a careful evaluation of the protein Delta-BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi)[72] search using as queries the CBM41 modules from *Thermotoga maritima* GH13_14 pullulanase (UniProt: O33840, residues S19-P120),[69] and *Streptococcus pneumoniae* GH13_12 glycogen-degrading enzyme SpuA (UniProt: Q97SQ7, two tandem modules CBM41-1 and CBM41-2 spanning residues D135-P240 and Q241-Y350, respectively).[3] Since it was evident from previous structural studies[6,25,69] that there are two versions of CBM41 with regard to the position in the sequence of the three key binding tryptophan (or aromatic) residues, i.e. the characteristic CBM41 binding motif, sequences were selected in an effort to include all relevant representatives of both versions including also those with rare substitutions of the three tryptophan positions.

After a collection of 2,448 CBM41 modules (not all presently classified in CAZy) originating from 1,666 protein sequences, a preliminary analysis was performed indicating that all CBM41s can be divided into nine groups (Fig. 1 and Table SI). It was thus decided to select CBM41s that would represent mainly the groups 1-6, i.e. those with conserved aromatic residues. Since the groups 1 and 5 contained a larger number of CBM41s in comparison with the groups 2, 3, 4 and 6, the set was composed of 100 CBM41s from each of the former two groups (i.e. 200 sequences) and 50 CBM41s from each of the latter four groups (i.e. also 200 sequences) (Table SI). These sequences were selected with regard to sequence differences and taxonomy within a given group in order to cover also eventual subgroups. All four eukaryotic representatives classified in the CAZy database (10 October 2016) were included. This selection has resulted, however, in finding four CBM41 copies that may be considered as "irregular" CBM41s, because they do not conform to the pattern of aromatic residues believed to be characteristic of the group to which they are assigned, but have a different aromatic substitution (Table SI). Concerning the groups 7-9 that were recognised as always lacking all three important aromatic residues, but exhibiting similarities to CBM41 within the remaining parts of their sequences, all 39 representatives from our original BLAST search and present in CAZy were used in the comparison (Table SI).

All full-length enzyme and protein sequences containing the CBM41 (Table SI) were retrieved from the UniProt knowledge database (http://www.uniprot.org/)[73] and/or from GenBank (https://www.ncbi.nlm.nih.gov/genbank/).[74] In every sequence the CBM41 boundaries were defined based on previous structural studies,[3,6,25,70,71] information available in the Pfam database (http://pfam.xfam.org/),[75] and data obtained from the BLAST searches described above. These boundaries were used to define the limits of the CBM sequences making up the input file for their alignment.

Sequence comparison and evolutionary relationships

The alignment was performed for the entire set of 443 CBM41 sequences originating from 284 amylolytic enzymes and hypothetical proteins using the program Clustal-Omega with default parameters (http://www.ebi.ac.uk/Tools/msa/clustalo/).[76] A subtle manual tuning was done in order to maximize similarities taking into account especially the residues identified as functionally important in family CBM41.[6,71] The evolutionary tree was calculated from the final alignment as a Phylip-tree type using the neighbour-joining clustering[77] and the bootstrapping procedure[78] (the number of bootstrap trials used was 1,000) implemented in the Clustal-X package.[79] The tree was displayed with the program iTOL (http://itol.embl.de/).[80]

Tertiary structure comparison

Three-dimensional structures of CBM41 were retrieved from the Protein Data Bank (PDB; http://www.rcsb.org/pdb/),[81] i.e., structures of the module from pullulanases of *Thermotoga maritima* (PDB code: 2J73),[6] *Klebsiella pneumoniae* (PDB code: 2FHB)[25] and *Klebsiella oxytoca* (PDB code: 2YOC)[70] as well as of the two tandem modules from the *Streptococcus pyogenes* surface-anchored pullulanase PulA (PDB code: 2J43)[3] and *Streptococcus pneumoniae* glycogen-degrading enzyme SpuA (PDB code: 2J44).[71] All the CBM41 structural models were created using the Phyre-2 server (http://www.sbg.bio.ic.ac.uk/phyre2/),[82] while structures were superimposed using the program MultiProt (http://bioinfo3d.cs.tau.ac.il/MultiProt/)[83] and displayed with the WebLab ViewerLite programme.

RESULTS AND DISCUSSION

CBM41 occurrence and sequence comparison

The present study delivers the *in silico* analysis of SBDs classified in the CAZy database as the family CBM41. Based on a preliminary search comprising data from both CAZy[1] and BLAST[72] and yielding 2,448 potential CBM41 sequences from 1,666 proteins, 443 CBMs originating from 284 proteins (Table SI) were selected for a detailed *in silico* analysis.

With regard to occurrence throughout the taxonomy spectrum, CBM41 has been found to be almost exclusively of bacterial origin, i.e. of all CBM41s collected, consistent with CAZy,[1] only four have originated from *Eucarya* and none from *Archaea*. Interestingly, all four eukaryotic CBM41s were of algal origin, either green algae – Chlorophyta (3 examples;

Bathycoccus prasinos, UniProt: K8EZ72; *Ostreococcus lucimarinus*, UniProt: A4RRA2; *Ostreococcus tauri*; UniProt: Q6PYZ0) or red algae – Rhodophyta (1 example; *Cyanidioschyzon merolae*, UniProt: M1VKA6).

Sequence comparison of the selected CBM41 modules, along with inspection of the available three-dimensional structures of the family CBM41 representatives belonging to different GH13 subfamilies of pullulanases, i.e. structures of the module from the subfamily GH13_13 *Klebsiella pneumoniae* pullulanase,[25] of the module from the subfamily GH13_14 *Thermotoga maritima* pullulanase[6] and of the tandem modules from the subfamily GH13_12 glycogen degrading enzyme SpuA of *Streptococcus pneumoniae,*[71] showed that, in most cases, there are two versions of arrangement in the amino acid sequence of the three crucial aromatic residues involved in maltooligosaccharide binding (Fig. 1). Both versions consist of a pair of tryptophan residues − or, in general, indispensable aromatic residues − separated in the sequence by only one residue (Trp-X-Trp), the former aromatic residue contributing a hydrogen bond involving the indole nitrogen and the latter providing a stacking interaction.[6,71] The difference between the two arrangements involves the position of the third tryptophan, which appears invariant and provides a stacking interaction. In one arrangement – seen, for example, in both modules from SpuA and also in the one from the *Klebsiella* pullulanase – the third tryptophan residue is located 10-11 residues after the second tryptophan from the above-mentioned pair (the *Streptococcus/Klebsiella* type), whereas in the other arrangement – evident, for example, in the *Thermotoga* module – this third tryptophan residue is placed about 30 residues toward the C-terminus from the position of the second tryptophan (i.e. the *Thermotoga* type). It is, however, worth mentioning that the third tryptophan coincides in space in the two versions, i.e., their side-chains are superimposable on each other in known three-dimensional structures.[6] There are, however, sequences where some or all of these aromatic residues are missing.

With respect to the existence of the two above-mentioned aromatic binding patterns (motifs), the alignment of CBM41s has allowed us to divide all sequences into 9 groups (Table SI) as follows: (i) groups 1-6 – CBM41 modules with the CBM41 motif of the three aromatic residues conserved within the group; and (ii) groups 7, 8 and 9 – the probable "non-binders" that are lacking at least one or even all the three aromatic residues from the CBM41 motif – or, if they are able to perform binding, other residues in their sequences have to be involved. Note that within the groups 1-6, CBM41s with an "irregular" aromatic motif may exist, that do not conform to the pattern for any one group. These are rather rare (only four CBM41s in the group 5 in Table SI). Their sequences are closely related to the "templates" from group 5, but they have a conservative substitution of one of the first two aromatic residues of the motif.

The sequences of groups 1-4 (the *Streptococcus/Klebsiella* type) as well as 5 and 6 (the *Thermotoga* type) possess three aromatic residues with approximately 10 and 30 residues, respectively, between the 2nd and the 3rd aromatic residue in the CBM41 binding motif. The sequences we classify in groups 1-6 possess the three aromatic residues believed to be important in carbohydrate binding. They are thought to represent starch- or glycogen-binding domains from the family CBM41, this conclusion being based on the invariant presence within each group of three aromatic residues forming the CBM41 characteristic binding motif (Fig. 1), identified in their crystal structures[3,6,25,70,71] and especially of complexes with maltooligosaccharides.[6,71] Note that although in the first two positions, a subtle variation is allowed, the third binding position is strictly occupied by a tryptophan residue (Fig. S1). The sequences that appear to contain three potential binding residues have been divided into six groups according to these variations and differences in the location of the third tryptophan residue.

Consideration of the Trp-X-Trp motif allows us to distinguish the sequences belonging to groups 1-4 from each other. In group 1 the pattern is invariantly Trp-X-Trp. In group 2 we see Trp-X-Phe, in group 3 Tyr-X-Trp and in group 4 His-X-Trp. Similarly, groups 5 and 6 can be distinguished from one another using this motif. In group 5 the motif usually is Trp-X-Trp, with a small minority of sequences (four of the sequences examined here) having Trp-X-Phe or His-X-Trp, but classified as belonging to group 5 because of close similarities along the length of their sequences with other group 5 members. In group 6, by contrast, the motif is Tyr-X-Trp.

Thus, of the first two aromatic binding residues separated from each other by only one non-conserved residue, the nature of the second aromatic residue seems to be more strictly conserved, i.e. of the six groups, only in group 2 is the tryptophan replaced by phenylalanine, whereas in the first position, the tryptophan is replaced by tyrosine in groups 3 and 6 or histidine in group 4 (Fig. 1). It should be noted, however, that in a small number of group 5 sequences the first tryptophan may be substituted by tyrosine or histidine. Group 4 is thus the only major group with the three aromatic binding residues conserved where one of these residues is histidine (Fig. S1). Interestingly, of the four CBM41s originating from *Eucarya*, three CBM41s from green algae are present in group 4. The fourth eukaryotic CBM41 from red algae belongs to the group 5 (Table SI).

With regard to the first “binding” position occupied in group 4 by a histidine, it can be pointed out that a hydrogen bond provider may be expected in that position, as evidenced, e.g., by structural details of CBM41 from *Thermotoga maritima* pullulanase.[6] The histidine represents thus a logical possibility. Interestingly, in groups 1-6, which may be the only “real binders” of all groups 1-9 described in the present study, only a histidine, in addition to tryptophan and tyrosine, has been found (Fig. S1). This strongly suggests a preference for a residue with an aromatic ring and H-bonding capability at that position, but for direct interaction with carbohydrate, it is the H-bonding capacity that is important.[3,6,71] Currently, however, the possibility that other H-bond donors might function just as well as the aromatic ones cannot simply be ruled out.

There are several other residues that play a functional role in CBM41. Thus in the *Thermotoga* type (groups 5 and 6), S35, E37, K76 and D81 were determined to be involved in various hydrogen bond contacts to carbohydrates.[6] Sequence comparison (Fig. S1) shows that, whereas both S35 and E37 are quite variable, the K76 and more convincingly the D81 may belong to very well conserved positions not only in the groups 5 and 6, but also in a substantial number of CBM41s belonging to the groups 1-4 (the *Streptococcus/Klebsiella* type).

With regard to groups 7-9, there is no experimental evidence in the literature on whether or not these CBM41s are able to bind α-glucans. Therefore, in accordance with their alignment (Fig. S1), the CBM41s classified into the groups 7-9 likely represent subfamilies of CBM41 “non-binders”. Alternatively other (aromatic) residues may have adopted binding roles. Indeed the present *in silico* analysis revealed that several positions (occupied mostly by aromatic residues), in addition to the three essential ones, are more-or-less conserved (Fig. S1). These are H13, Y14, Y21, W24, F43, F49, F83, W94 and F103 (numbering according to in *Thermotoga maritima* pullulanase[6]). It is of interest that at least some of them are invariant among the CBM41 representatives of the groups 7-9 (H13, Y14, Y21 and W94) or are substituted by another aromatic residue or, due to uncertainty in the alignment, have an aromatic counterpart in the adjacent position(s) (Fig. S1). Since these additional positions containing aromatic residues are neither totally conserved, nor have any functional role ascribed based on solved CBM41 structures,[6,71] it is hypothesised that these positions may represent a relict from a primordial CBM ancestor before the current CBM41s specialized during evolution. Certainly most of these residues are not close, as deduced from comparison with the *Thermotoga maritima* CBM41 module, to the carbohydrate binding site and may simply play a role in stabilising the folding of the module. It has already been revealed that some, but probably not all, CBM families considered to be SBDs have related folds, and form either something like a CBM clan[20] or exhibit sequence similarities sharing a mode of α-glucan binding characteristic of a CBM.[7-14]

Domain arrangement

Figure 2 shows the domain arrangement of selected representative proteins bearing the CBM41s that belong to the individual groups identified here. The main catalytic domain of each of these proteins is from the α-amylase family GH13[4,5] and mostly with the specificity of a pullulanase. This is supported by solved tertiary structures[3,6,25,70,71] as well as by affiliation of these sequences to GH13 subfamilies 12, 13 and 14 which represent the pullulanase enzyme specificity (EC 3.2.1.41).[66] Many sequences studied here share a domain arrangement observed for bacterial pullulanases from *Klebsiella pneumoniae*[25,84] and *Thermotoga maritima*[6,69] and their hypothetical homologue from green algae *Ostreococcus tauri* (UniProt: Q6PYZ0), representing the groups 3, 5 and 4, respectively, regardless of their pullulanase GH13 subfamily (Fig. 2). Simultaneously, there are many pullulanases and/or their putative counterparts that possess more than one CBM41 copy. These multiple copies may or may not belong to the same group, e.g., the glycogen degrading enzyme SpuA from *Streptococcus pneumoniae*[71] and the putative pullulanase PulA from *Streptococcus pyogenes*[3] each have two copies of CBM41, but in the former case both copies belong to group 1, whereas in the latter case the first copy belongs to group 1 and the second copy to group 2 (Fig. 2).

Another example of domain organization can be illustrated by α-amylase-pullulanases consisting of two independent catalytic domains – an α-amylase of subfamily GH13_32 and a pullulanase of subfamily GH13_14 – bearing, in the case of the enzyme from *Bifidobacterium breve* (UniProt: F9Y041), two CBM41 copies belonging to groups 6 and 5 in that order (Fig. 2). It should be pointed out that there are also enzymes having four copies of CBM41 (Table SI), either with all four copies within a single group or with two copies from one group and two copies from another, e.g., putative pullulanases with two catalytic domains from *Streptococcus gallolyticus* (UniProt: F5WVK9 – group 1) and *Bacillus akibai* (UniProt: W4QRN9 – groups 5 and 1 in that order).

Concerning the CBM41s from group 7, they are usuall*y* found in tandem succeeded by their counterparts from group 4 (Table SI), both copies being located C-terminally with regard to position of the catalytic domain of enzymes from the subfamily GH13_13 (not shown). On the other hand, the CBM41s from groups 8 and 9 belong to putative pullulanases from the subfamily GH13_14 with a domain arrangement similar to that seen in the *Thermotoga* *maritima* pullulanase,[6,69] with the CBM48 being positioned closer to the C-terminus of CBM41 (cf. Fig. 2). Finally, some CBM41s are present within shorter proteins (~200-300 residues) seemingly without any catalytic domain (Table SI).

Although the CBM41s may clearly sometimes exist in tandem (especially in groups 1 and 2), this is not always true (Fig. 2). Nevertheless, at least in the case of glycogen degrading enzyme from *Streptococcus pneumoniae* and pullulanase from *Streptococcus pyogenes* (and their homologues), the two CBM41 modules of each enzyme molecule form dyads that act together to help in binding the substrate.[3,71] There are examples of existence in tandems for SBDs from other families, e.g., CBM25 and CBM26 both in GH13 α-amylases[8,11,36,85-87] as well as CBM20 in GH77 4-α-glucanotransferases (disproportionating enzymes DPE2),[7,88,89] but for none of these a three-dimensional structure was solved. This co-operative action in a dyad cannot be the case for enzymes with only one CBM41. If the CBM41 is a part of a family GH13 enzyme, however, available evidence suggests there is always a CBM48 directly preceding the catalytic GH13 domain (Fig. 2). In the pullulanase from *Klebsiella pneumoniae*, an additional domain with a CBM-like fold is present between the CBM41 (from the group 3) and the CBM48.[25] Both known CBMs of this pullulanase, i.e. CBM41 and CBM48, seem to be far enough apart in the three-dimensional structure that they are unlikely to interact with one another. The possibility of an enzyme acting as a dimer, however, should be taken into account. As described for members of the neopullulanase subfamily possessing the CBM34,[90-92] in these enzymes their CBM34 of one molecule interacts with the active site and bound substrate of the other molecule of the dimer.[24,93-95] This is exactly what was observed in the crystal structure of *Klebsiella pneumoniae* pullulanase after the substrate was bound.[25] Moreover, a very close homologue, the pullulanase from *Klebsiella aerogenes* is believed to act as a monomer in solution[96] so here there is no indication of CBM41s acting as a dyad.

With regard to catalytic domain in terms of the affiliation to the main α-amylase family GH13[4,5] and taking into account the information on available tertiary structures,[3,6,25,70,71] we see that the groups 1, 3 and 5 are represented typically in the subfamilies GH13_12 (*Streptococcus* glycogen-degrading enzyme), GH13_13 (*Klebsiella* pullulanases) and GH13_14 (*Thermotoga* pullulanase), respectively (Fig. 2; cf. Table SI).

Structural comparison

For structural comparison of CBM41s belonging to the individual groups 1-9 defined in the present study (Table SI), representative tertiary structures were either retrieved from PDB[81]

(CBM41s from groups 1, 2, 3 and 5) or modelled via the Phyre-2 server[82] (CBM41s from groups 4, 6, 7, 8 and 9). It is worth mentioning here that all models were produced allowing the server to choose the best templates. In each case, i.e. for the representatives of the groups 4, 6, 7, 8 and 9, the modelling resulted in 10 meaningful models (with 100% confidence) based solely on CBM41 templates that span at least 94-95% of the modelled sequence. Among all the relevant CBM41 templates, the one from *Thermotoga maritima* pullulanase (PDB code: 2J73)[6] exhibited the highest sequence identity with the representative of each group; therefore the models based on this template were selected for further analysis. However, since the preliminary tertiary structure inspection indicated inappropriate orientation of one of the three potential binding residues in the CBM41 from *Ostreococcus tauri*, the model constructed according to the CBM41 template from *Klebsiella pneumoniae* pullulanase (PDB code: 2FHF)[25] was definitively selected although there was a lower sequence identity between the template and modelled sequences (Fig. 3). Using this latter template then gave a model with orientation of binding residues very similar to those found in the models for the other CBM41s that were based on the *Thermotoga maritima* pullulanase.

Despite some differences, the overall fold adopts a typical CBM41-like β-sandwich structure[17] identified previously in CBM modules of typical family GH13 pullulanases[3,6,25,70,71] (Fig. 3). These modules have the CBM41-characteristic motif of three aromatic residues involved in binding α-glucans seen in the representatives of groups 1-6 (Fig. 3a-f). It has already been demonstrated for CBM41 structures from *Klebsiella* and *Thermotoga* pullulanases[6] that all the three functional aromatic residues are fully superimposable, i.e. they are structurally conserved despite the fact that, at the sequence level, the third aromatic binding residue (tryptophan) of groups 1-4 does not correspond to the equivalent tryptophan residue of groups 5 and 6 (Fig. 3). Variations seem to be allowed – at least according to our current knowledge – only at the first two positions of the motif since the third position is always occupied by a tryptophan. Group 4 with the histidine residue in the motif His-X-Trp (Fig. 3d) is presented for CBM41 for the first time here.

As far as CBM41s from the "non-binders" groups 7, 8 and 9 are concerned, all of them have been modelled using the CBM41 of the pullulanase from *Thermotoga maritima*[6] as the best template. Since these sequences have already been classified into the family CBM41 in CAZy[1] (cf. Table SI), they, as expected, exhibit a fold typical for this SBD family (Fig. 3g-i) despite the fact that they do not possess the three aromatic residues recognised as the CBM41 binding motif. Each representative structure from the groups 7-9 gives a convincing result for superimposition with the structure of its template, i.e. the overlays cover in each case almost the entire sequence length with a meaningful value of root-mean square deviation (RMSD). This enables identification even of residues corresponding with the binding motif (Fig. 3j-l), which were arginine, glycine and lysine for the group 7 (Fig. 3j), serine, lysine and none as the third residue for the group 8 (Fig. 3k), and valine, serine and glutamine for the group 9 (Fig. 3l). But for all these, additional experimental work focusing on biochemical characterization is necessary to either prove or eliminate their carbohydrate-binding ability, since at present they all are only in the CAZy database[1] as putative proteins released from genome sequencing projects. It is, however, worth mentioning that there are several other positions with aromatic residues in their sequences (Fig. 1) that are well conserved throughout these groups (Fig. S1) that could be checked by, e.g. site-directed mutagenesis in order to verify whether or not they have a binding role.

It is of note that the binding residues – or residues corresponding to them – were in each case deduced by inspection of structure-based sequence alignment provided by the Phyre2 server[82] and supported by superimposition of a structural model with its template performed by MultiProt server.[83] The results from both servers suggest that all the three potential binding aromatic residues in the CBM41 models of the groups 4 and 6 (Fig. 3d,f) have been identified correctly and that there are most probably no aromatic residues in the CBM41 models of the groups 7, 8 and 9 that would correspond with binding residues from the known CBM41 structures (Fig. 3j-l).

Evolutionary relationships

The phylogenetic tree of the members of the SBD family CBM41, divided into nine groups according to their CBM41-characteristic binding motif, is illustrated in Figure 4 and demonstrates the overall evolutionary picture, i.e. clustering of all 443 CBM41s from the present study (Table SI). The same tree with necessary details on all the individual 443 CBM41s is shown in Figure S2.

It is evident that there are two main parts in the evolutionary tree: (i) the “binders” groups 1-4 that includes also the “non-binders” group 7; and (ii) the “binders” group 5 and 6 covering also the “non-binders” groups 8 and 9. Thus the two parts of the tree reflect the basic difference between the CBM41-characteristic binding motif, i.e. the distance between the second aromatic binding residue and the third tryptophan being either ~10 or ~30 amino acid residues (cf. Figs. 2 and 3).

With regard to the part of the tree comprising the groups with the shorter distance between the two above-mentioned residues, it is clear that groups 1 and 2, originating mostly from the genus *Streptococcus,* are closely related as are groups 3 and 4 having predominantly the *Klebsiella*, *Vibrio* and *Streptomyces* origin, the last- mentioned cluster comprising also the three CBM41s of eukaryotic origin (green algae). The group 7 with all of its members being from the genus *Streptomyces* is found to be most closely related with their taxonomical counterparts of group 3 (Fig. S2).

As far as the second part of the tree is concerned, involving the groups with the longer distance in the CBM41 binding motif, the groups 8 and 9 (again all from the genus *Streptococcus*) are located on adjacent branches sharing a cluster with streptococcal representatives from group 5. Taxonomy in this part of the tree is further reflected especially for bacilli from the genera *Bacillus* (group 5) and *Lactobacillus* (groups 5 and 6) as well as for actinobacteria from *Bifidobacterium* and *Gardnerella* that can be found in two different clusters representing both groups 5 and 6 separately. The only eukaryotic representative CBM41 from the red alga (*Cyanidioschyzon merolae*) is positioned in a small taxonomically variable cluster of the group 5 together with CBM41s from *Listeriaceae*, *Helcococcus*, *Spirochaeta*, *Paenibacillus* and *Mobiluncus* (Fig. S2). The four "irregular" CBM41s of group 5, which originate from Firmicutes (Table SI) and have a conservative substitution (i.e. an aromatic residue for another aromatic residue) in the CBM41 binding motif, are found in the tree also in agreement with their taxonomy. Since the groups 5 and 6 cover a wider spectrum of organisms, it is reasonable to see more variability among their clusters in comparison with groups 1-4 that comprise a substantially narrower taxonomical sample (Table SI).

Overall, it can be concluded that as originally observed for SBDs from the family CBM20 from various amylolytic GH families[23] and recently confirmed also for this SBD present in non-amylolytic starch-degrading lytic polysaccharide monooxygenases,[50,97-100] taxonomy seems to be more respected in the evolution of CBMs than the specificity of the enzyme to which a given CBM sequence is attached.

Conclusions

The present *in silico* study allows the separation of the family CBM41 into two basic subdivisions: (i) the so-called *Streptococcus*/*Klebsiella*-type; and (ii) the so-called *Thermotoga*-type. These two subdivisions can be distinguished from each other by a characteristic sequence motif of the three essential aromatics, i.e. "W-W-~10aa-W" for the former type and "W-W-~30aa-W" for the latter type. Of the three crucial aromatic residues constituting one carbohydrate-binding site in the CBM41 β-sandwich fold, the tryptophans in the first and second positions of the motif can be substituted by other aromatic residues, such as phenylalanine or tyrosine, or even histidine. This has resulted in the observation of six different CBM41 carbohydrate-binding groups – four in the *Streptococcus*/*Klebsiella*-type and two in the *Thermotoga*-type. All the members of these six groups, reflecting mostly differences in taxonomy, should retain the ability to bind carbohydrate (α-glucans). Three additional groups have been proposed that obviously lack the entire aromatic motif, but for these the use of residues from remaining parts of their sequence for binding remains a possibility. The analysis of domain arrangement indicates that the SBD of the family CBM41, if present as a module in an enzyme, is found only in enzymes with the family GH13 α-amylase-type catalytic TIM-barrel domain. Moreover, those enzymes are most typically pullulanases of the subfamilies GH13_12, 13 and 14, and the CBM41 is often present in these enzymes together with the SBD of family CBM48. In the overall evolutionary picture of the family CBM41, taxonomy, i.e. the evolution of species, seems to be most reflected in both subdivisions – among the individual groups and also within a group.

ACKNOWLEDGEMENTS

SJ thanks for financial support the Slovak Grant Agency VEGA for the grants No. 2/0150/14 and 2/0146/17, and the Slovak Research and Development Agency for the contract No. LPP-0417-09. BS thanks for financial support from the Danish Council for Independent Research | Natural Sciences, grant No. 6108-00476. Karol Blesák is acknowledged for being engaged during his PhD-studies in the beginning of this project. Special thanks to Marian Salaj for his computerized help with preliminary working with large data sets. All authors disclose no conflict of interest.

AUTHOR CONTRIBUTIONS

SJ designed the study, contributed to collecting data, analysed and interpreted results, prepared figures and wrote the manuscript; KM collected data, analysed results and prepared figures; BS contributed to designing the study, interpreting results and writing the manuscript; EAM contributed to interpreting results and writing the manuscript.

REFERENCES

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

2 Lammerts van Bueren A, Finn R, Ausio J, Boraston AB. α-Glucan recognition by a new family of carbohydrate-binding modules found primarily in bacterial pathogens. Biochemistry 2004;43:15633–15642.

3 van Bueren AL, Higgins M, Wang D, Burke RD, Boraston AB. Identification and structural basis of binding to host lung glycogen by streptococcal virulence factors. Nat Struct Mol Biol 2007;14:76–84.

4 MacGregor EA, Janecek S, Svensson B. Relationship of sequence and structure to specificity in the α-amylase family of enzymes. Biochim Biophys Acta 2001;1546:1–20.

5 Janecek S, Svensson B, Macgregor EA. α-Amylase: an enzyme specificity found in various families of glycoside hydrolases. Cell Mol Life Sci 2014;71:1149–1170.

6 van Bueren AL, Boraston AB. The structural basis of α-glucan recognition by a family 41 carbohydrate-binding module from *Thermotoga maritima*. J Mol Biol 2007;365:555–560.

7 Janecek S, Svensson B, MacGregor EA. Structural and evolutionary aspects of two families of non-catalytic domains present in starch and glycogen binding proteins from microbes, plants and animals. Enzyme Microb Technol 2011;49:429–440.

8 Majzlova K, Janecek S. Two structurally related starch-binding domain families CBM25 and CBM26. Biologia 2014;69:1087–1096.

9 Svensson B, Jespersen H, Sierks MR, MacGregor EA. Sequence homology between putative raw-starch binding domains from different starch-degrading enzymes. Biochem J 1989;264:309–311.

10 Rodriguez-Sanoja R, Oviedo N, Sanchez S. Microbial starch-binding domain. Curr Opin Microbiol 2005;8:260–267.

11 Machovic M, Janecek S. Starch-binding domains in the post-genome era. Cell Mol Life Sci 2006;63:2710–2724.

12 Christiansen C, Abou Hachem M, Janecek S, Viksø-Nielsen A, Blennow A, Svensson B. The carbohydrate-binding module family 20 - diversity, structure, and function. FEBS J 2009;276:5006–5029.

13 Guillen D, Sanchez S, Rodriguez-Sanoja R. Carbohydrate-binding domains: multiplicity of biological roles. Appl Microbiol Biotechnol 2010;85:1241–1249.

14 Carvalho CC, Phan NN, Chen Y, Reilly PJ. Carbohydrate-binding module tribes. Biopolymers 2015;103:203–214.

15 Sorimachi K, Le Gal-Coëffet MF, Williamson G, Archer DB, Williamson MP. Solution structure of the granular starch binding domain of *Aspergillus niger* glucoamylase bound to β-cyclodextrin. Structure 1997;5:647–661.

16 Southall SM, Simpson PJ, Gilbert HJ, Williamson G, Williamson MP. The starch-binding domain from glucoamylase disrupts the structure of starch. FEBS Lett 1999;447:58–60.

17 Bozic N, Loncar N, Sokarda Slavic M, Vujcic Z. Raw starch degrading α-amylases: an unsolved riddle. Amylase 2017;1:12–25.

18 Boraston AB, Bolam DN, Gilbert HJ, Davies GJ. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. Biochem J 2004;382:769–781.

19 Hashimoto H. Recent structural studies of carbohydrate-binding modules. Cell Mol Life Sci 2006;63:2954-2967.

20 Machovic M, Svensson B, MacGregor EA, Janecek S. A new clan of CBM families based on bioinformatics of starch-binding domains from families CBM20 and CBM21. FEBS J 2005;272:5497–5513.

21 Liu YN, Lai YT, Chou WI, Chang MD, Lyu PC. Solution structure of family 21 carbohydrate-binding module from *Rhizopus oryzae* glucoamylase. Biochem J 2007;403:21–30.

22 Jiang TY, Ci YP, Chou WI, Lee YC, Sun YJ, Chou WY, Li KM, Chang MD. Two unique ligand-binding clamps of *Rhizopus oryzae* starch binding domain for helical structure disruption of amylose. PLoS One 2012;7:e41131.

23 Janecek S, Sevcik J. The evolution of starch-binding domain. FEBS Lett 1999;456:119–125.

24 Abe A, Tonozuka T, Sakano Y, Kamitori S. Complex structures of *Thermoactinomyces vulgaris* R-47 α-amylase 1 with malto-oligosaccharides demonstrate the role of domain N acting as a starch-binding domain. J Mol Biol 2004;335:811–822.

25 Mikami B, Iwamoto H, Malle D, Yoon HJ, Demirkan-Sarikaya E, Mezaki Y, Katsuya Y. Crystal structure of pullulanase: evidence for parallel binding of oligosaccharides in the active site. J Mol Biol 2006;359:690–707.

26 Mikkelsen R, Suszkiewicz K, Blennow A. A novel type carbohydrate-binding module identified in α-glucan, water dikinases is specific for regulated plastidial starch metabolism. Biochemistry 2006;45:4674–4682.

27 Tung JY, Chang MD, Chou WI, Liu YY, Yeh YH, Chang FY, Lin SC, Qiu ZL, Sun YJ. Crystal structures of the starch-binding domain from *Rhizopus oryzae* glucoamylase reveal a polysaccharide-binding path. Biochem J 2008;416:27–36.

28 Wayllace NZ, Valdez HA, Ugalde RA, Busi MV, Gomez-Casati DF. The starch-binding capacity of the noncatalytic SBD2 region and the interaction between the N- and C-terminal domains are involved in the modulation of the activity of starch synthase III from *Arabidopsis thaliana*. FEBS J 2010;277:428–440.

29 Glaring MA, Baumann MJ, Abou Hachem M, Nakai H, Nakai N, Santelia D, Sigurskjold BW, Zeeman SC, Blennow A, Svensson B. Starch-binding domains in the CBM45 family - low-affinity domains from glucan, water dikinase and α-amylase involved in plastidial starch metabolism. FEBS J 2011;278:1175–1185.

30 Møller MS, Abou Hachem M, Svensson B, Henriksen A. Structure of the starch-debranching enzyme barley limit dextrinase reveals homology of the N-terminal domain to CBM21. Acta Crystallogr Sect F Struct Biol Cryst Commun 2012;68:1008–1012.

31 Orzechowski S, Grabowska A, Sitnicka D, Siminska J, Felus M, Dudkiewicz M, Fudali S, Sobczak M. Analysis of the expression, subcellular and tissue localisation of phosphoglucan, water dikinase (PWD/GWD3) in *Solanum tuberosum* L.: a bioinformatics approach for the comparative analysis of two α-glucan, water dikinases (GWDs) from *Solanum tuberosum* L. Acta Physiol Plant 2013;35:483–500.

32 Xu J, Ren F, Huang CH, Zheng Y, Zhen J, Sun H, Ko TP, He M, Chen CC, Chan HC, Guo RT, Song H, Ma Y. Functional and structural studies of pullulanase from *Anoxybacillus* sp. LM18-11. Proteins 2014;82:1685–1693.

33 Penninga D, van der Veen BA, Knegtel RM, van Hijum SA, Rozeboom HJ, Kalk KH, Dijkstra BW, Dijkhuizen L. The raw starch binding domain of cyclodextrin glycosyltransferase from *Bacillus circulans* strain 251. J Biol Chem 1996;271:32777–32784.

34 Mikami B, Adachi M, Kage T, Sarikaya E, Nanmori T, Shinke R, Utsumi S. Structure of raw starch-digesting *Bacillus cereus* β-amylase complexed with maltose. Biochemistry 1999;38:7050–7061.

35 Oyama T, Kusunoki M, Kishimoto Y, Takasaki Y, Nitta Y. Crystal structure of β-amylase from *Bacillus cereus* var. mycoides at 2.2 Å resolution. J Biochem 1999;125:1120–1130.

36 Sumitani J, Tottori T, Kawaguchi T, Arai M. New type of starch-binding domain: the direct repeat motif in the C-terminal region of *Bacillus* sp. no. 195 α-amylase contributes to starch binding and raw starch degrading. Biochem J 2000;350:477–484.

37 Boraston AB, Healey M, Klassen J, Ficko-Blean E, Lammerts van Bueren A, Law V. A structural and functional analysis of α-glucan recognition by family 25 and 26 carbohydrate-binding modules reveals a conserved mode of starch recognition. J Biol Chem 2006;281:587–598.

38 Bott R, Saldajeno M, Cuevas W, Ward D, Scheffers M, Aehle W, Karkehabadi S, Sandgren M, Hansson H. Three-dimensional structure of an intact glycoside hydrolase family 15 glucoamylase from *Hypocrea jecorina*. Biochemistry 2008;47:5746–5754.

39 Rodriguez-Sanoja R, Oviedo N, Escalante L, Ruiz B, Sanchez S. A single residue mutation abolishes attachment of the CBM26 starch-binding domain from *Lactobacillus amylovorus* α-amylase. J Ind Microbiol Biotechnol 2009;36:341–346.

40 Peng H, Zheng Y, Chen M, Wang Y, Xiao Y, Gao Y. A starch-binding domain identified in α-amylase (AmyP) represents a new family of carbohydrate-binding modules that contribute to enzymatic hydrolysis of soluble starch. FEBS Lett 2014;588:1161–1167.

41 Valk V, van der Kaaij RM, Dijkhuizen L. The evolutionary origin and possible functional roles of FNIII domains in two *Microbacterium aurum* B8.A granular starch degrading enzymes, and in other carbohydrate acting enzymes. Amylase 2017;1:1–11.

42 Koropatkin NM, Smith TJ. SusG: a unique cell-membrane-associated α-amylase from a prominent human gut symbiont targets complex starch molecules. Structure 2010;18:200–215.

43 Valk V, Lammerts van Bueren A, van der Kaaij RM, Dijkhuizen L. Carbohydrate-binding module 74 is a novel starch-binding domain associated with large and multidomain α-amylase enzymes. FEBS J 2016;283:2354–2368.

44 Janeček Š, Svensson B. Amylolytic glycoside hydrolases. Cell Mol Life Sci. 2016;73:2601–2602.

45 Foley MH, Cockburn DW, Koropatkin NM. The Sus operon: a model system for starch uptake by the human gut Bacteroidetes. Cell Mol Life Sci. 2016;73:2603–2617.

46 Møller MS, Henriksen A, Svensson B. Structure and function of α-glucan debranching enzymes. Cell Mol Life Sci. 2016;73:2619–2641.

47 Mahlow S, Orzechowski S, Fettke J. Starch phosphorylation: insights and perspectives. Cell Mol Life Sci. 2016;73:2753–2764.

48 Emanuelle S, Brewer MK, Meekins DA, Gentry MS. Unique carbohydrate binding platforms employed by the glucan phosphatases. ell Mol Life Sci. 2016;73:2765–2778.

49 Pfister B, Zeeman SC. Formation of starch in plant cells. Cell Mol Life Sci. 2016;73:2781–2807.

50 Vu VV, Marletta MA. Starch-degrading polysaccharide monooxygenases. Cell Mol Life Sci 2016;73:2809–2819.

51 Palopoli N, Busi MV, Fornasari MS, Gomez-Casati D, Ugalde R, Parisi G. Starch-synthase III family encodes a tandem of three starch-binding domains. Proteins 2006;65:27–31.

52 Valdez HA, Busi MV, Wayllace NZ, Parisi G, Ugalde RA, Gomez-Casati DF. Role of the N-terminal starch-binding domains in the kinetic properties of starch synthase III from *Arabidopsis thaliana*. Biochemistry 2008;47:3026–3032.

53 Seung D, Soyk S, Coiro M, Maier BA, Eicke S, Zeeman SC. Protein targeting to starch is required for localising granule-bound starch synthase to starch granules and for normal amylose synthesis in *Arabidopsis*. PLoS Biol 2015;13:e1002080.

54 Vander Kooi CW, Taylor AO, Pace RM, Meekins DA, Guo HF, Kim Y, Gentry MS. Structural basis for the glucan phosphatase activity of Starch Excess4. Proc Natl Acad Sci USA 2010;107:15379–15384.

55 Meekins DA, Raththagala M, Husodo S, White CJ, Guo HF, Kötting O, Vander Kooi CW, Gentry MS. Phosphoglucan-bound structure of starch phosphatase Starch Excess4 reveals the mechanism for C6 specificity. Proc Natl Acad Sci USA 2014;111:7272–7277.

56 Minassian BA, Ianzano L, Meloche M, Andermann E, Rouleau GA, Delgado-Escueta AV, Scherer SW. Mutation spectrum and predicted function of laforin in Lafora's progressive myoclonus epilepsy. Neurology 2000;55:341–346.

57 Gentry MS, Romá-Mateo C, Sanz P. Laforin, a protein with many faces: glucan phosphatase, adapter protein, et alii. FEBS J 2013;280:525–537.

58 Raththagala M, Brewer MK, Parker MW, Sherwood AR, Wong BK, Hsu S, Bridges TM, Paasch BC, Hellman LM, Husodo S, Meekins DA, Taylor AO, Turner BD, Auger KD, Dukhande VV, Chakravarthy S, Sanz P, Woods VL Jr, Li S, Vander Kooi CW, Gentry MS. Structural mechanism of laforin function in glycogen dephosphorylation and Lafora disease. Mol Cell 2015;57:261–272.

59 Sankhala RS, Koksal AC, Ho L, Nitschke F, Minassian BA, Cingolani G. Dimeric quaternary structure of human laforin. J Biol Chem 2015;290:4552–4559.

60 Janecek S. A motif of a microbial starch-binding domain found in human genethonin. Bioinformatics 2002;18:1534–1537.

61 Jiang S, Heller B, Tagliabracci VS, Zhai L, Irimia JM, DePaoli-Roach AA, Wells CD, Skurat AV, Roach PJ. Starch binding domain-containing protein 1/genethonin 1 is a novel participant in glycogen metabolism. J Biol Chem 2010;285:34960–34971.

62 Sun T, Yi H, Yang C, Kishnani PS, Sun B. Starch binding domain-containing protein 1 plays a dominant role in glycogen transport to lysosomes in liver. J Biol Chem 2016;291:16479–16484.

63 Polekhina G, Gupta A, van Denderen BJ, Feil SC, Kemp BE, Stapleton D, Parker MW. Structural basis for glycogen recognition by AMP-activated protein kinase. Structure 2005;13:1453–1462.

64 Hardie DG. AMP-activated protein kinase: an energy sensor that regulates all aspects of cell function. Genes Dev 2011;25:1895–1908.

65 Avila-Castaneda A, Gutiérrez-Granados N, Ruiz-Gayosso A, Sosa-Peinado A, Martínez-Barajas E, Coello P. Structural and functional basis for starch binding in the SnRK1 subunits AKINβ2 and AKINβγ. Front Plant Sci 2014:5:199.

66 Emanuelle S, Hossain MI, Moller IE, Pedersen HL, van de Meene AM, Doblin MS, Koay A, Oakhill JS, Scott JW, Willats WG, Kemp BE, Bacic A, Gooley PR, Stapleton DI. SnRK1 from *Arabidopsis thaliana* is an atypical AMPK. Plant J 2015;82:183–192.

67 Vu VV, Beeson WT, Span EA, Farquhar ER, Marletta MA. A family of starch-active polysaccharide monooxygenases. Proc Natl Acad Sci USA 2014;111:13822–13827.

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

69 Bibel M, Brettl C, Gosslar U, Kriegshäuser G, Liebl W. Isolation and analysis of genes for amylolytic enzymes of the hyperthermophilic bacterium *Thermotoga maritima*. FEMS Microbiol Lett 1998;158:9–15.

70 East A, Mechaly AE, Huysmans GH, Bernarde C, Tello-Manigne D, Nadeau N, Pugsley AP, Buschiazzo A, Alzari PM, Bond PJ, Francetic O. Structural basis of pullulanase membrane binding and secretion revealed by X-ray crystallography, molecular dynamics and biochemical analysis. Structure 2016;24:92–104.

71 Lammerts van Bueren A, Ficko-Blean E, Pluvinage B, Hehemann JH, Higgins MA, Deng L, Ogunniyi AD, Stroeher UH, El Warry N, Burke RD, Czjzek M, Paton JC, Vocadlo DJ, Boraston AB. The conformation and function of a multimodular glycogen-degrading pneumococcal virulence factor. Structure 2011;19:640–651.

72 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.

73 UniProt Consortium.. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res 2014;42:D191-D198.

74 Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res 2014;42:D32–D37.

75 Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M (2014) Pfam: the protein families database. Nucleic Acids Res 42, D222–D230.

76 Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 2011;7:539.

77 Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4:406–425.

78 Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 1985;39:783–791.

79 Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins D. Clustal W and Clustal X version 2.0. Bioinformatics 2007;23:2947–2948.

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

81 Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.

82 Kelley LA, Sternberg MJ. Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc 2009;4:363–371.

83 Shatsky M, Nussinov R, Wolfson HJ. A method for simultaneous alignment of multiple protein structures. Proteins 2004;56:143–156.

84 Katsuragi N, Takizawa N, Murooka Y. Entire nucleotide sequence of the pullulanase gene of *Klebsiella aerogenes* W70. J Bacteriol 1987;169:2301–2306.

85 Rodriguez Sanoja R, Morlon-Guyot J, Jore J, Pintado J, Juge N, Guyot JP. Comparative characterization of complete and truncated forms of *Lactobacillus amylovorus* α-amylase and role of the C-terminal direct repeats in raw-starch binding. Appl Environ Microbiol 2000;66:3350–3356.

86 Rodriguez-Sanoja R, Ruiz B, Guyot JP, Sanchez S. 2005. Starch-binding domain affects catalysis in two *Lactobacillus* α-amylases. Appl Environ Microbiol 2005;71:297–302.

87 Yamaguchi R, Tokunaga H, Ishibashi M, Arakawa T, Tokunaga M. Salt-dependent thermo-reversible α-amylase: cloning and characterization of halophilic α-amylase from moderately halophilic bacterium, *Kocuria varians*. Appl Microbiol Biotechnol 2011;89:673–684.

88 Lloyd JR, Blennow A, Burhenne K, Kossmann J. Repression of a novel isoform of disproportionating enzyme (stDPE2) in potato leads to inhibition of starch degradation in leaves but not tubers stored at low temperature, Plant Physiol 2004;134:1347–1354.

89 Kuchtova A, Janecek S. *In silico* analysis of family GH77 with focus on amylomaltases from borreliae and disproportionating enzymes DPE2 from plants and bacteria. Biochim Biophys Acta 2015;1854:1260–1268.

90 Park KH, Kim TJ, Cheong TK, Kim JW, Oh BH, Svensson B. Structure, specificity and function of cyclomaltodextrinase, a multispecific enzyme of the α-amylase family. Biochim Biophys Acta 2000;1478:165–185.

91 Oslancova A, Janecek S. Oligo-1,6-glucosidase and neopullulanase enzyme subfamilies from the α-amylase family defined by the fifth conserved sequence region. Cell Mol Life Sci 2002;59:1945–1959.

92 Kuchtova A, Janecek S. Domain evolution in enzymes of the neopullulanase subfamily. Microbiology 2016;162:2099–2115.

93 Kim JS, Cha SS, Kim HJ, Kim TJ, Ha NC, Oh ST, Cho HS, Cho MJ, Kim MJ, Lee HS, Kim JW, Choi KY, Park KH, Oh BH. Crystal structure of a maltogenic amylase provides insights into a catalytic versatility. J Biol Chem 1999;274:26279–26286.

94 Lee HS, Kim MS, Cho HS, Kim JI, Kim TJ, Choi JH, Park C, Lee HS, Oh BH, Park KH. Cyclomaltodextrinase, neopullulanase, and maltogenic amylase are nearly indistinguishable from each other. J Biol Chem 2002;277:21891–21897.

95 Hondoh H, Kuriki T, Matsuura Y. Three-dimensional structure and substrate binding of *Bacillus stearothermophilus* neopullulanase. J Mol Biol 2003;326:177–188.

96 Takizawa N, Murooka Y. Cloning of the pullulanase gene and overproduction of pullulanase in *Escherichia coli* and *Klebsiella aerogenes*. Appl Environ Microbiol 1985;49:294–298.

97 Nekiunaite L, Isaksen T, Vaaje-Kolstad G, Abou Hachem M. Fungal lytic polysaccharide monooxygenases bind starch and β-cyclodextrin similarly to amylolytic hydrolases. FEBS Lett 2016;590:2737–2747.

98 Lo Leggio L, Simmons TJ, Poulsen JC, Frandsen KE, Hemsworth GR, Stringer MA, von Freiesleben P, Tovborg M, Johansen KS, De Maria L, Harris PV, Soong CL, Dupree P, Tryfona T, Lenfant N, Henrissat B, Davies GJ, Walton PH. Structure and boosting activity of a starch-degrading lytic polysaccharide monooxygenase. Nat Commun 2015;6:5961.

99 Nekiunaite L, Arntzen MO, Svensson B, Vaaje-Kolstad G, Abou Hachem M. Lytic polysaccharide monooxygenases and other oxidative enzymes are abundantly secreted by *Aspergillus nidulans* grown on different starches. Biotechnol Biofuels 2016;9:187.

100 Frandsen KE, Poulsen JC, Tovborg M, Johansen KS, Lo Leggio L. Learning from oligosaccharide soaks of crystals of an AA13 lytic polysaccharide monooxygenase: crystal packing, ligand binding and active-site disorder. Acta Crystallogr D Struct Biol 2017;73:64–76.

Figure Legends

Figure 1

Sequence alignment of selected CBM41 representatives from groups 1-9. Colour code for the aromatic residues: W – yellow, F – red; Y – green; H – blue. The positions occupied by aromatic residues, forming the characteristic CBM41 pattern/motif responsible for α-glucan binding, are signified by asterisks above the alignment. Hashtags indicate the nine additional positions with partially conserved aromatic residues, discussed in the main text, potentially playing a functional role, especially in CBM41s from the groups 7, 8 and 9. The label of the CBM41 consists of the name of the organism and the UniProt accession number succeeded by a digit (if any) in the case where there are multiple copies of CBM41 within a single protein.

Figure 2

Domain arrangement of selected CBM41 representatives from groups 1-6. The position and size of all domains in individual representatives were extracted with help of literature[3,6,25,71] and Pfam database[75], and should correlate with those in real proteins (UniProt accession number is given in parenthesis). GH13 denotes the α-amylase family; its subfamily[68], if known, is indicated. The other CBM families of SBDs, if present, are also shown.

Figure 3

Structural comparison of selected CBM41 representatives from groups 1-9. (a-f) Groups 1-6; (g-i) groups 7-9. Sources of the proteins: (a) Glycogen degrading enzyme SpuA from *Streptococcus pneumoniae* (PDB code: 2J44) – Q97SQ7_1; (b) pullulanase PulA from *Streptococcus pyogenes* (PDB code: 2J43) – Q8KLP1_2; (c) pullulanase from *Klebsiella pneumoniae* (PDB code: 2FHB) – P08711; (d) putative pullulanase from *Ostreococcus tauri* (model) – Q6PYZ0; (e) pullulanase from *Thermotoga maritima* (PDB code: 2J73) – O33840; (f) α-amylase-pullulanase from *Bifidobacterium breve* (model) – F9Y041_1; (g) putative family GH13 member from *Streptomyces flavogriseus* (model) – E8WFD0_1; (h) putative family GH13 member from *Streptococcus oralis* (model) – F2QDA0; (i) putative family GH13 member from *Streptococcus pneumoniae* (model) – B2IPZ0. For structural models, the CBM41 of the pullulanase from *Thermotoga maritima* (group 5; PDB: 2J73) was the best template for representatives of groups 6 (30% sequence identity and 98% sequence coverage), 7 (33% and 97%), 8 (28% and 98%) and 9 (25% and 98%), whereas the CBM41 of the pullulanase from *Klebsiella pneumoniae* (group 3; PDB: 2FHF) was used for modelling the representative of the group 4 (16% sequence identity and 94% sequence coverage). The confidence of all models was 100% in each case meaning that the sequence of the model and its template are true homologues and that the overall fold is modelled at high accuracy. For the groups 1-6 (a-f), the CBM41 characteristic pattern/motifs of the three aromatic residues responsible for α-glucan binding are shown, the first residue being in the middle, the second residue being at the bottom and the third one at the top. (j-l) CBM41 modelled structures from the groups 7, 8 and 9 coloured olive, green and walnut, respectively, superimposed with their template – CBM41 from *Thermotoga maritima* pullulanase (coloured blue) with bound maltotetraose (coloured by elements). (j) Alignment size: 81 C_α atoms, RMSD: 0.63 Å; (k) Alignment size: 85 C_α atoms, RMSD: 0.73 Å; (l) Alignment size: 84 C_α atoms, RMSD: 0.83 Å. It is clear that in the CBM41s from the groups 7-9, the three aromatic residues from the CBM41 characteristic binding pattern/motif do not have their aromatic counterparts.

Figure 4

Evolutionary tree of the SBD family CBM41. The tree is based on the alignment of all 443 CBM41 sequences derived from 284 proteins (Fig. S1). The nine groups proposed are highlighted by different colours, the groups being also indicated by relevant digits near their clusters. For the sake of simplicity, only the branches are shown. The same tree with all the leaves described is presented in Figure S2. Details concerning all CBM41 sequences compared in the tree as well as their colouring are given in Table SI.

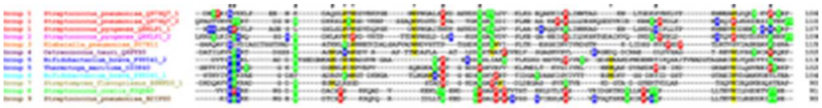

Figure 1. Sequence alignment of selected CBM41 representatives from groups 1-9.

34x4mm (300 x 300 DPI)

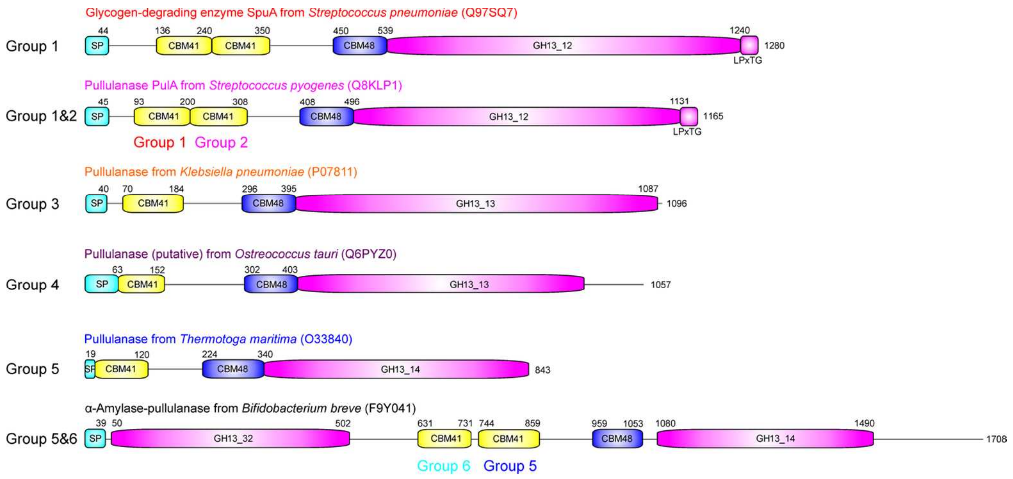

Figure 2. Domain arrangement of selected CBM41 representatives from groups 1-6.

128x60mm (300 x 300 DPI)

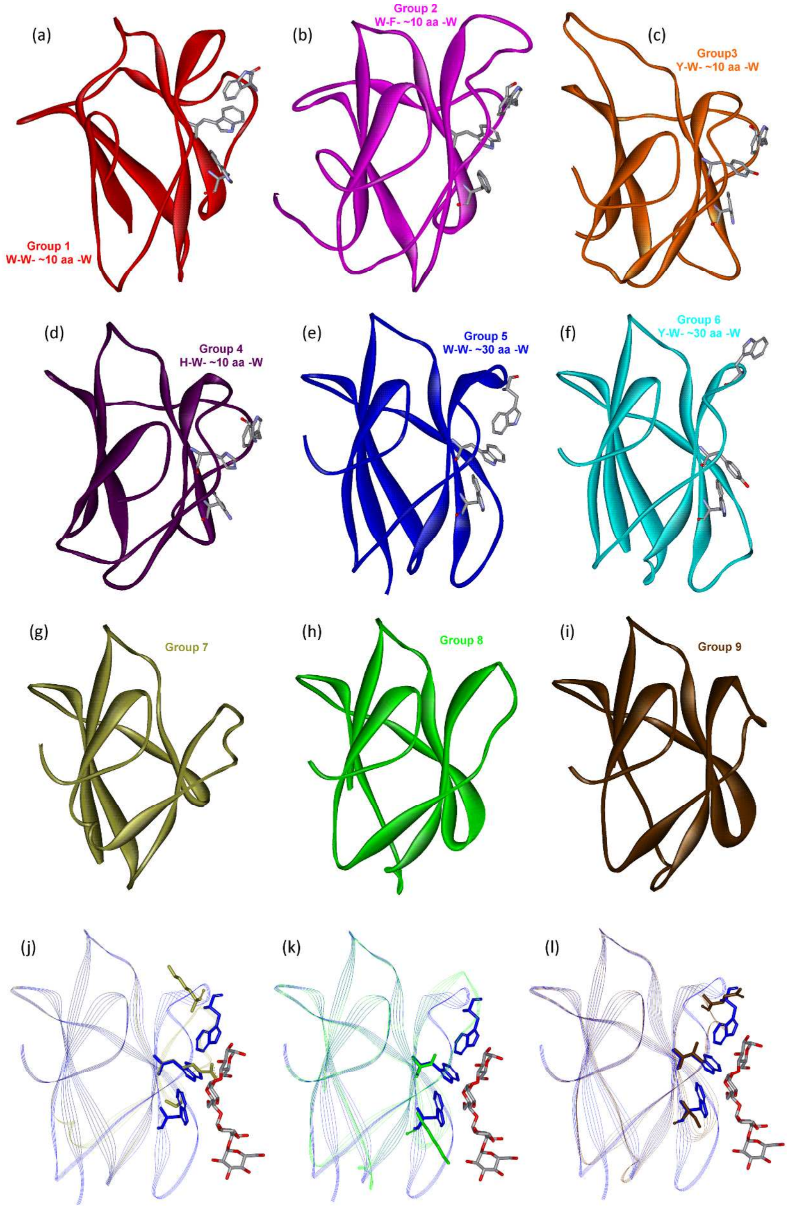

Figure 3. Structural comparison of selected CBM41 representatives from groups 1-9.

257x388mm (300 x 300 DPI)

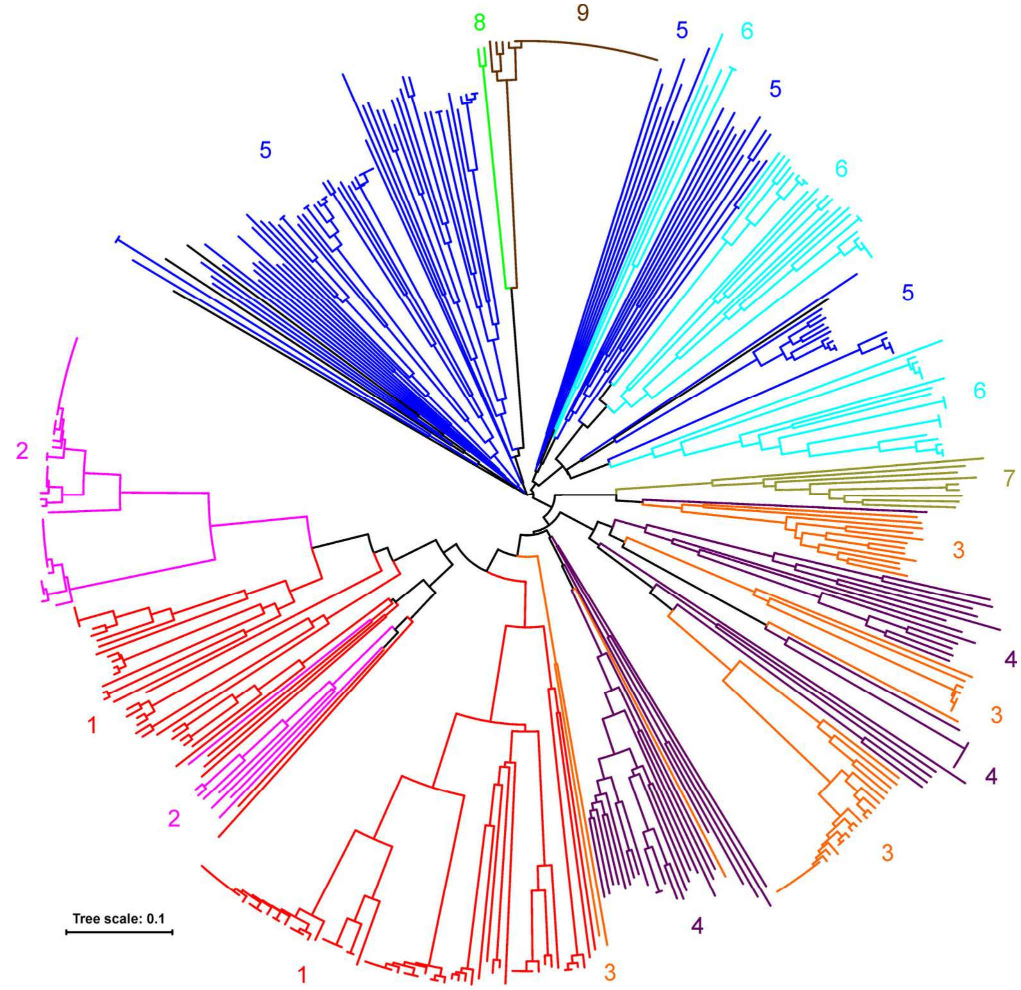

Figure 4. Evolutionary tree of the SBD family CBM41.

177x173mm (300 x 300 DPI)