the FEBS Journal

A new clan of CBM families based on bioinformatics of starch-binding domains from families CBM20 and CBM21

Martin Machovič[1], Birte Svensson[2], E. Ann MacGregor[3] and Štefan Janeček[1]

[1] Institute of Molecular Biology, Slovak Academy of Sciences, Bratislava, Slovakia
[2] Biochemistry and Nutrition Group, BioCentrum-DTU, Technical University of Denmark, Kgs. Lyngby, Denmark
[3] 2 Nicklaus Green, Livingston, West Lothian, UK

Keywords
carbohydrate-binding module; evolutionary tree; glycoside hydrolase family; sequence alignment; starch-binding domain

Correspondence
Š. Janeček, Institute of Molecular Biology, member of the Centre of Excellence for Molecular Medicine, Slovak Academy of Sciences, Dúbravská cesta 21, SK-84551 Bratislava 45, Slovakia
Fax: +421 25930 7416
Tel: +421 25930 7420
E-mail: stefan.janecek@savba.sk

(Received 27 May 2005, revised 13 July 2005, accepted 30 August 2005)

doi:10.1111/j.1742-4658.2005.04942.x

Approximately 10% of amylolytic enzymes are able to bind and degrade raw starch. Usually a distinct domain, the starch-binding domain (SBD), is responsible for this property. These domains have been classified into families of carbohydrate-binding modules (CBM). At present, there are six SBD families: CBM20, CBM21, CBM25, CBM26, CBM34, and CBM41. This work is concentrated on CBM20 and CBM21. The CBM20 module was believed to be located almost exclusively at the C-terminal end of various amylases. The CBM21 module was known as the N-terminally positioned SBD of *Rhizopus* glucoamylase. Nowadays many nonamylolytic proteins have been recognized as possessing sequence segments that exhibit similarities with the experimentally observed CBM20 and CBM21. These facts have stimulated interest in carrying out a rigorous bioinformatics analysis of the two CBM families. The present analysis showed that the original idea of the CBM20 module being at the C-terminus and the CBM21 module at the N-terminus of a protein should be modified. Although the CBM20 functionally important tryptophans were found to be substituted in several cases, these aromatics and the regions around them belong to the best conserved parts of the CBM20 module. They were therefore used as templates for revealing the corresponding regions in the CBM21 family. Secondary structure prediction together with fold recognition indicated that the CBM21 module structure should be similar to that of CBM20. The evolutionary tree based on a common alignment of sequences of both modules showed that the CBM21 SBDs from α-amylases and glucoamylases are the closest relatives to the CBM20 counterparts, with the CBM20 modules from the glycoside hydrolase family GH13 amylopullulanases being possible candidates for the intermediate between the two CBM families.

Amylolytic enzymes are multidomain proteins. The three best known are α-amylase (EC 3.2.1.1), β-amylase (EC 3.2.1.2) and glucoamylase (EC 3.2.1.3) [1,2], which differ structurally and functionally from each other. In the sequence-based classification CAZy [3] of glycoside hydrolases (GH) they belong to the independent families GH13, GH14 and GH15, respectively, which have no mutual sequence similarities.

Family GH13 contains enzymes with about 30 different enzyme specificities [4] and forms, together with GH70 and GH77, the clan GH-H [5]. Unrelated α-amylases and amylolytic enzymes with sequence similarities to such α-amylases were grouped into family GH57 [6], while some amylolytic enzymes are also found in family GH31 [7]. The amylolytic enzymes belonging to the clan GH-H (families GH13, GH70,

Abbreviations
CBM, carbohydrate-binding module; CGTase, cyclodextrin glucanotransferase; GH, glycoside hydrolase family; SBD, starch-binding domain.

and GH77) are distinctly different from those found in families GH14, GH15, GH31, and GH57 in terms of amino acid sequences and three-dimensional structures. Moreover, these families employ different reaction mechanisms and catalytic machineries. The members of GH13 (α-amylases), GH14 (β-amylases) and a GH31 xylosidase adopt different $(\beta/\alpha)_8$-barrel folds for the catalytic domain [8–10], while the catalytic domain in GH15 (glucoamylases) is a helical $(\alpha/\alpha)_6$-barrel fold [11]. The structure of a GH57 4-α-glucanotransferase was recently determined as a $(\beta/\alpha)_7$-barrel [12]. As far as the reaction mechanism is concerned, α-amylases and related enzymes (clan GH-H), as well as the enzymes from GH31 and GH57, employ a retaining mechanism, whereas β-amylases (GH14) and glucoamylases (GH15) are inverting enzymes [13,14].

Approximately 10% of all amylolytic enzymes possess a distinct domain enabling binding and degradation of raw starch. Certain amylolytic enzymes have this capacity without the presence of a specialized functional domain [15–17], but these are few. One example is the barley α-amylase that binds to raw starch at a surface binding site on the catalytic domain. This has been demonstrated by mutational analysis [15] and the site is seen as two critically oriented tryptophan residues in the crystal structure of the complex with acarbose [18]. A second surface site was recently discovered in the C-terminal domain, which seems unique to barley α-amylase 1 [19]. Mutational analysis of this site demonstrated a binding role [20]. Based on their sequences the starch-binding domains (SBD) have also been classified into families of carbohydrate-binding modules (CBM) [21]. At present, there are six SBD families in CAZy (recently reviewed in [22]): CBM20, CBM21, CBM25, CBM26, CBM34, and CBM41 [23–31].

The present work focuses on SBD families CBM20 and CBM21. The CBM20 module is ≈ 90–130 residues long and has been studied most intensively. It is located in most cases at the C-terminus of amylolytic enzymes from families GH13, GH14, and GH15 [23,24]. The three-dimensional structure of the isolated SBD alone has been determined by NMR as well as by X-ray crystallography of enzymes that contain this SBD [32–38]. The CBM20 module consists of seven β-strand segments forming an open-sided distorted β-barrel. Several aromatics, especially the well-conserved Trp and Tyr residues, were proposed to be essential for the function of the SBD [23], and these were confirmed to participate in two raw starch-binding sites of the module [39–43]. It has been demonstrated that, if fused to another protein, this SBD independently retains its function even when the target protein is not an amylase [44–48]. On the other hand, there is a lack of information on structure–function relationships of the CBM21 module. The length in this case varies in the range ≈ 90–140. The CBM21 module is well known as the N-terminally positioned SBD of *Rhizopus oryzae* glucoamylase [49]. Recently several nonamylolytic proteins (especially as deduced from sequenced genomes) were recognized to possess amino acid sequence stretches that exhibit unambiguous similarities with the experimentally observed SBDs of CBM20 and CBM21, e.g. protein phosphatases (EC 3.1.3.16).[50], laforin [51], and genethonin-1 [52]. These observations strongly motivated interest in carrying out a rigorous bioinformatics analysis of the two CBM families.

A structural relationship between the C-terminally positioned (CBM20) and the N-terminally positioned (CBM21) SBDs was suggested more than 15 years ago, based on sequence alignments [23]. We therefore, in the first step, analyzed the sequences of both families separately, taking into account the above-mentioned lack of structure–function information concerning CBM21. This was followed by attempts to identify the CBM20 sequence of structural features in the sequences of CBM21, aimed at revealing amino acid residues that correspond with each other in the two families. Finally, a sequence alignment was made that served for calculation of the common CBM20-CBM21 evolutionary tree. This provides a basis for the joining of the two CBMs into a common clan.

Results and Discussion

Location of SBD modules in CBM20 and CBM21

With regard to the location of the SBD in the polypeptide chain, analysis of recent sequences showed that the original idea [23,24] of the CBM20 module being at the C-terminus and the CBM21 module at the N-terminus of a protein, should be modified (Fig. 1). Thus, the division into C-terminal and N-terminal SBDs seems to hold for the SBDs possessing the established function of raw starch-binding, while the other proteins (nonamylases), exhibiting only the sequence motif features of CBM20 or CBM21, do not necessarily obey this rule. It is worth mentioning that the real starch-binding function could be ascribed only to α-amylase (GH13), β-amylase (GH14), glucoamylase (GH15), maltooligosaccharide-producing amylases (GH13), cyclodextrin glucanotransferase [CGTase, (EC 2.4.1.19)] (GH13), and acarviose transferase (GH13) that altogether constitute less than 30% of the sequences, i.e., more than 60% in the family CBM20 and only about 10% in CBM21.

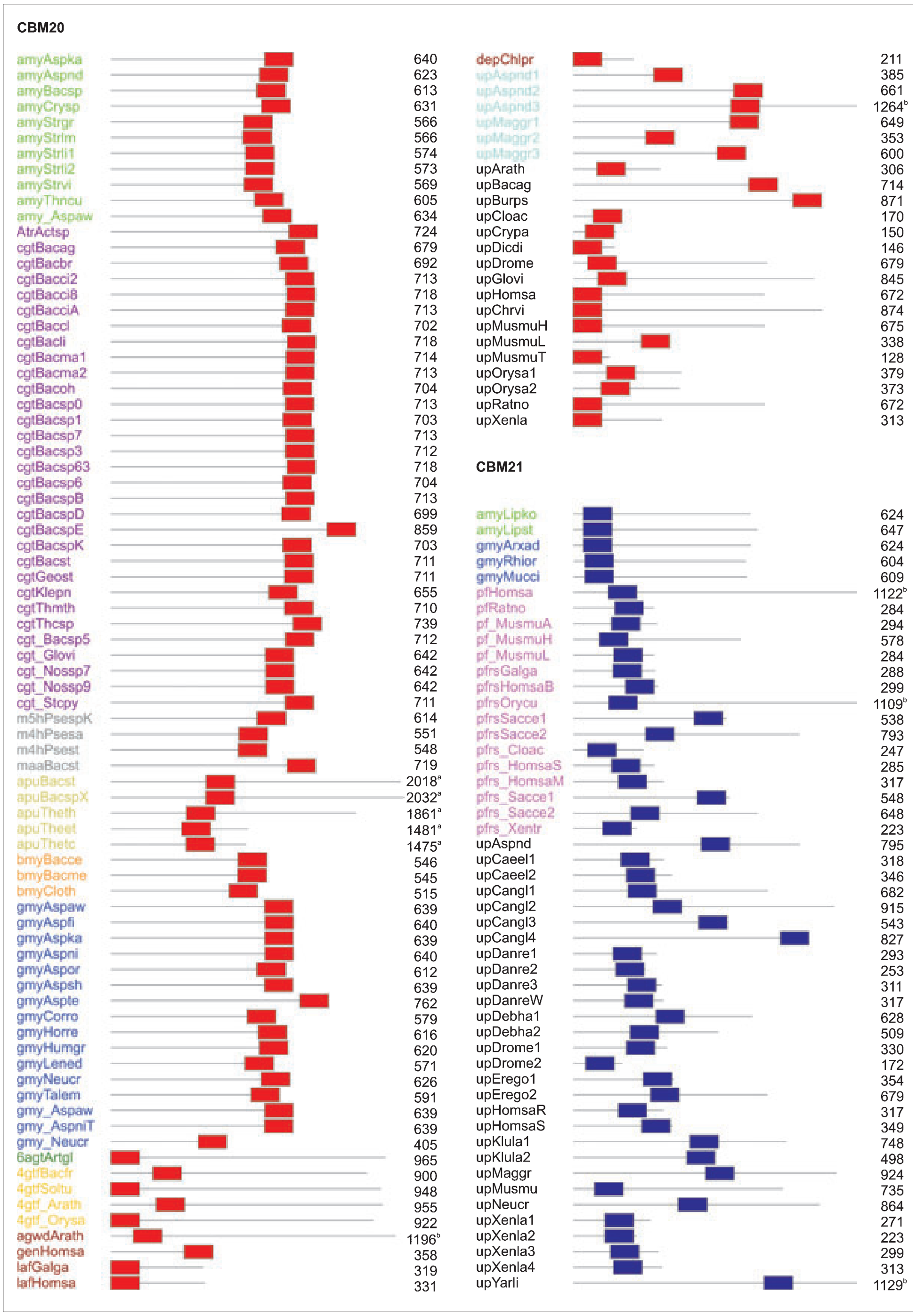

Fig. 1. Position of the CBM20 and CBM21 modules in the amino acid sequences. For the proteins without (^{a}) or (^{b}), these are the total lengths of the proteins and the black lines are drawn to scale to represent protein lengths. For the proteins with (^{a}) and (^{b}), 1000 residues from the N-terminus are deleted and shown, respectively. For example, for apuBacst (2018^{a}), the protein is 2018 residues long, but only the last 1018 are shown; and for agwdArath (1196^{b}), the protein is 1196 residues long, but only the first 1000 from the N-terminal end are shown. For protein identification, see Table 1.

There are several other glycoside hydrolases containing the CBM20 module, e.g. amylopullulanase (GH13), 6-α-glucosyltransferase (GH31), and 4-α-glucanotransferase (GH77), for which a real starch-binding function has not been demonstrated up to now. These CBM20 modules are positioned inside the polypeptide chain (amylopullulanases) or at the N-terminal end (6-α-glucosyltransferase and 4-α-glucanotransferases). Interestingly, α-glucan water dikinase, a starch phosphorylating enzyme from *Arabidopsis thaliana*, contains a CBM20 module near the N-terminal end of the protein. The N-terminal location is also seen in the case of the majority of unknown proteins of eukaryotic origin with a recognized CBM20 module (Fig. 1). At present it is not possible to decide the real function of CBM20 in these proteins, with a single remarkable exception, laforin [51], the protein product of the Lafora type of epilepsy gene, which was proven experimentally to bind starch with its CBM20 module [53,54].

The situation in CBM21 is more complicated, because microbial amylolytic enzymes represent only 10% of the sequences in this family. A substantial number of the remaining CBM21 members are eukaryotic protein phosphatases and/or their regulatory subunits. Interestingly, the regulatory subunit, called the glycogen-targeting G subunit, was shown to direct the protein phosphatase to glycogen [55]. Because these proteins were shown to also contain a binding site for glycogen phosphorylase, they, albeit indirectly, also play a role in glycogen metabolism [56]. At present the majority of the CBM21 family modules belong to unknown proteins of various origins. As far as the location of the SBD is concerned, this module is clearly neither positioned N-terminally (except for the amylases) nor exclusively at or near the C-terminal end of the protein (Fig. 1). Thus CBM20 and CBM21 can no longer be considered as exclusively C- and N-terminally positioned, respectively. It should be noted, however, that up until now CBM21 has been found only in eukaryotes (Table 1).

Sequence analysis

Detailed analysis of amino acid sequences of the SBDs revealed that CBM20 has no invariant residues, whereas CBM21 has a single invariant Lys34 (*Rhizopus oryzae* glucoamylase numbering) (Fig. 2; the complete alignment is not shown).

Originally 11 consensus residues were shown for a small number of CBM20 sequences [23]. Their structural arrangements in the motifs from the representatives of bacteria and fungi are illustrated in Fig. 3. As the number of sequences increased, a few (about 2%) substitutions were found at these positions [24]. At present even the functionally important tryptophans, Trp643, Trp689 of binding site 1 (Fig. 3; *Bacillus circulans* strain 251 CGTase numbering, i.e., the Trp616 and Trp662 after removing the 27-residue long signal peptide), are not absolutely conserved. While the former tryptophan is missing in only one case (CBM20 motif of the CGTase from *Streptococcus pyogenes*), the latter varies more often (Fig. 2). Interestingly Trp689 is substituted in all three putative CGTases from cyanobacteria (*Gloeobacter violaceous*, *Nostoc* sp. PCC7120 and PCC9229), all five amylopullulanases, one glucoamylase (*Hormoconis resinae*), two 4-α-glucanotransferases (*Arabidopsis thaliana* and rice), and two unknown proteins (upAspni3, upMaggr2) (Fig. 2). However, no sequence lacks both of these signature tryptophans. The region around Trp643 (residues LGxW) is the best conserved part of the entire CBM20 motif. As far as the remaining consensus residues are concerned, these are best conserved in amylolytic enzymes, with the exception of amylopullulanases, which, however, do contain the equivalent of Lys678 (Fig. 2) associated with binding site 1 (Fig. 3; *B. circulans* CGTase numbering).

Besides the consensus residues, the present analysis identified the position equivalent to Phe618 (*B. circulans* CGTase numbering, i.e., the Phe591 after removing the 27-residue long signal peptide) as highly conserved (87.5%). This phenylalanine is present not only in the amylolytic enzymes, but also in the animal SBDs as found in laforin and genethonin-1 (Fig. 2). The lack of this residue in the three putative CGTases of cyanobacteria and the CGTase from *S. pyogenes* is remarkable. These sequences are unusual in other ways, however, in that the cyanobacterial CGTases lack the equivalent of Trp689 (Trp662 without the signal peptide), while the *S. pyogenes* CGTase lacks the essential tryptophan from the region LGxW.

At present it is not possible to say more about the real function of SBDs from the cyanobacterial CGTases included in the present analysis. The CGTases from *Gloeobacter violaceus* and *Nostoc* sp. PCC7120 were identified in the complete genome sequences [57,58], while that from *Nostoc* sp. PCC9229 was cloned and expressed as a putative CGTase [59]. It seems that not all cyanobacteria must contain the putative CGTase gene, e.g. it is missing from the genome of *Synechocystis* sp. 6803 [60].

Despite numerous substitutions observed in the consensus positions (Fig. 2), the regions around these residues remain the best conserved segments of a SBD of CBM20 type. They were thus used as markers to reveal possible correspondence with CBM21 as well as to adjust CBM20 and CBM21 sequences to each other. Although the probable relatedness of the two SBD families was indicated more than 15 years ago [23], the lack of the three-dimensional structure of CBM21 makes it less straightforward to deduce whether or not the two CBM modules are related. It is remarkable,

Table 1. The enzymes and proteins containing the CBM20 and CBM21 modules. The abbreviation 'prot. phosp. reg. sub.' means the regulatory subunit of protein phosphatase. All sequences were retrieved from GenBank except for the cgtBacma2 (UniProt: P31835).

Abbreviation	Specificity	EC number	Source	GenBank	Length	Glycoside hydrolase family
CBM20						
(Bright green of Fig.2)						
amyAspka	α-amylase	3.2.1.1	*Aspergillus kawachi*	BAA22993	640	13
amyAspnd	α-amylase	3.2.1.1	*Aspergillus nidulans*	AAF17100	623	13
amyBacsp	α-amylase	3.2.1.1	*Bacillus* sp. TS-23	AAA63900	613	13
amyCrysp	α-amylase	3.2.1.1	*Cryptococcus* sp. S-2	BAA12010	631	13
amyStrgr	α-amylase	3.2.1.1	*Streptomyces griseus*	CAA40798	566	13
amyStrlm	α-amylase	3.2.1.1	*Streptomyces limosus*	AAA88554	566	13
amyStrli1	α-amylase	3.2.1.1	*Streptomyces lividans*	CAA73926	574	13
amyStrli2	α-amylase	3.2.1.1	*Streptomyces lividans*	CAB06622	573	13
amyStrvi	α-amylase	3.2.1.1	*Streptomyces violaceus*	AAB36561	569	13
amyThncu	α-amylase	3.2.1.1	*Thermomonospora curvata*	CAA41881	605	13
amy_Aspaw	α-amylase	n.d.	*Aspergillus awamori*	BAD06003	634	13
CBM20						
(Purple of Fig.2)						
atrActsp	acarviose transferase	2.4.1.19	*Actinoplanes* sp. 50/110	AAE37556	724	13
cgtBacag	CGTase	2.4.1.19	*Bacillus agaradhaerens*	AAP31242	679	13
cgtBacbr	CGTase	2.4.1.19	*Bacillus brevis*	AAB65420	692	13
cgtBacci2	CGTase	2.4.1.19	*Bacillus circulans* 251	CAA55023	713	13
cgtBacci8	CGTase	2.4.1.19	*Bacillus circulans* 8	CAA48401	718	13
cgtBacciA	CGTase	2.4.1.19	*Bacillus circulans* A11	AAG31622	713	13
cgtBaccl	CGTase	2.4.1.19	*Bacillus clarkii*	BAB91217	702	13
cgtBacli	CGTase	2.4.1.19	*Bacillus licheniformis*	CAA33763	718	13
cgtBacma1	CGTase	2.4.1.19	*Bacillus macerans*	AAA22298	714	13
cgtBacma2	CGTase	2.4.1.19	*Bacillus macerans*	P31835	713	13
cgtBacoh	CGTase	2.4.1.19	*Bacillus ohbensis*	BAA14289	704	13
cgtBacsp0	CGTase	2.4.1.19	*Bacillus* sp. 1011	AAA22308	713	13
cgtBacsp1	CGTase	2.4.1.19	*Bacillus* sp. 1-1	ALBSX1	703	13
cgtBacsp7	CGTase	2.4.1.19	*Bacillus* sp. 17-1	AAA22310	713	13
cgtBacsp3	CGTase	2.4.1.19	*Bacillus* sp. 38-2	AAA22309	712	13
cgtBacsp63	CGTase	2.4.1.19	*Bacillus* sp. 6.3.3	CAA46901	718	13
cgtBacsp6	CGTase	2.4.1.19	*Bacillus* sp. 633	BAA31539	704	13
cgtBacspB	CGTase	2.4.1.19	*Bacillus* sp. B1018	AAA22239	713	13
cgtBacspD	CGTase	2.4.1.19	*Bacillus* sp. DSM 5850	CAA01436	699	13
cgtBacspE	CGTase	2.4.1.19	*Bacillus* sp. E-1	Z34466	859	13
cgtBacspK	CGTase	2.4.1.19	*Bacillus* sp. KC201	BAA02380	703	13
cgtBacst	CGTase	2.4.1.19	*Bacillus stearothermophilus*	CAA41770	711	13
cgtGeost	CGTase	2.4.1.19	*Geobacillus stearothermophilus*	AAD00555	711	13
cgtKlepn	CGTase	2.4.1.19	*Klebsiella pneumonie*	AAA25059	655	13
cgtThmth	CGTase	2.4.1.19	*Thermoanaerobacter thermosulfurogenes*	AAB00845	710	13
cgtThcsp	CGTase	2.4.1.19	*Thermococcus* sp. B1001	BAA88217	739	13
cgt_Bacsp5	CGTase	n.d.	*Bacillus* sp. I-5	AAR32682	712	13
cgt_Glovi	CGTase	n.d.	*Gloeobacter violaceus*	BAC88314	642	13
cgt_Nossp7	CGTase	n.d.	*Nostoc* sp. PCC 7120	BAB77693	642	13
cgt_Nossp9	CGTase	n.d.	*Nostoc* sp. PCC 9229	AAM16154	642	13
cgt_Stcpy	CGTase	n.d.	*Streptococcus pyogenes*	AAK34149	711	13
(Grey of Fig. 2)						
m5hPsespK	maltopentaohydrolase	3.2.1.-	*Pseudomonas* sp. KO-8940	BAA01600	614	13
m4hPsesa	maltotetraohydrolase	3.2.1.60	*Pseudomonas saccharophila*	CAA34708	551	13
m4hPsest	maltotetraohydrolase	3.2.1.60	*Pseudomonas stutzeri*	AAA25707	548	13
maaBacst	maltogenic α-amylase	3.2.1.133	*Bacillus stearothermophilus*	AAA22233	719	13

Table 1. (Continued).

Abbreviation	Specificity	EC number	Source	GenBank	Length	Glycoside hydrolase family
(Dark yellow of Fig. 2)						
apuBacst	amylopullulanase	3.2.1.41	*Bacillus stearothermophilus*	AAG44799	2018	13
apuBacspX	amylopullulanase	3.2.1.41	*Bacillus* sp. XAL601	BAA05832	2032	13
apuTheth	amylopullulanase	3.2.1.41	*Thermoanaerobacter thermosulfurogenes*	AAB00841	1861	13
apuTheet	amylopullulanase	3.2.1.41	*Thermoanaerobacter ethanolicus*	AAA23201	1481	13
apuThetc	amylopullulanase	3.2.1.41	*Thermoanaerobacter thermohydrosulfuricus*	AAA23205	1475	13
(Red of Fig.2)						
bmyBacce	β-amylase	3.2.1.2	*Bacillus cereus*	BAA34650	546	14
bmyBacme	β-amylase	3.2.1.2	*Bacillus megaterium*	CAB61483	545	14
bmyCloth	β-amylase	3.2.1.2	*Clostridium thermosulfurogenes*	AAA23204	515	14
(Blue of Fig. 2)						
gmyAspaw	glucoamylase	3.2.1.3	*Aspergillus awamori*	AAB02927	639	15
gmyAspfi	glucoamylase	3.2.1.3	*Aspergillus ficuum*	AAT58037	640	15
gmyAspka	glucoamylase	3.2.1.3	*Aspergillus kawachi*	BAA00331	639	15
gmyAspni	glucoamylase	3.2.1.3	*Aspergillus niger*	AAB59296	640	15
gmyAspor	glucoamylase	3.2.1.3	*Aspergillus oryzae*	AAB20818	612	15
gmyAspsh	glucoamylase	3.2.1.3	*Aspergillus shirousami*	BAA01254	639	15
gmyAspte	glucoamylase	3.2.1.3	*Aspergillus tereus*	L15383	762	15
gmyCorro	glucoamylase	3.2.1.3	*Corticium rolfsii*	BAA08436	579	15
gmyHorre	glucoamylase	3.2.1.3	*Hormoconis resinae*	CAA47945	616	15
gmyHumgr	glucoamylase	3.2.1.3	*Humicola grisea*	AAA33386	620	15
gmyLened	glucoamylase	3.2.1.3	*Lentinula edodes*	AAF75523	571	15
gmyNeucr	glucoamylase	3.2.1.3	*Neurospora crassa*	AAE15056	626	15
gmyTalem	glucoamylase	3.2.1.3	*Talaromyces emersonii*	AAR61398	591	15
gmy_Aspaw	glucoamylase	n.d.	*Aspergillus awamori*	BAD06004	639	15
gmy_AspniT	glucoamylase	n.d.	*Aspergillus niger* T21	AAP04499	639	15
gmy_Neucr	glucoamylase	n.d.	*Neurospora crassa*	CAE75704	405	15
(Green of Fig. 2)						
6agtArtgl	6-α-glucosyltransferase	n.d.	*Arthrobacter globiformis*	BAD34980	965	31
(Yellow of Fig. 2)						
4agtBacfr	4-α-glucanotransferase	2.4.1.25	*Bacteroides fragilis*	BAD50570	900	77
4agtSoltu	4-α-glucanotransferase	2.4.1.25	*Solanum tuberosum*	AAR99599	948	77
4agt_Arath	4-α-glucanotransferase	n.d.	*Arabidopsis thaliana*	AAL91204	955	77
4agt_Orysa	4-α-glucanotransferase	n.d.	*Oryza sativa*	BAC22431	922	77
(Dark red of Fig. 2)						
agwdArath	α-glucan water dikinase	2.7.9.4	*Arabidopsis thaliana*	AY747068	1196	–
genHomsa	genethonin-1	–	*Homo sapiens*	AAH22301	358	–
lafGalga	laforin	–	*Gallus gallus*	CAG31547	319	–
lafHomsa	laforin	–	*Homo sapiens*	AAG18377	331	–
depChlpr	degreenig enhanced protein	–	*Chlorella protothecoides*	CAB42581	211	–
(Turquoise of Fig. 2)						
upAspnd1	unknown protein	–	*Aspergillus nidulans*	EAA62623	385	–
upAspnd2	unknown protein	–	*Aspergillus nidulans*	EAA61773	661	–
upAspnd3	unknown protein	–	*Aspergillus nidulans*	EAA64118	1264	–
upMaggr1	unknown protein	–	*Magnaporthe grisea*	XP_368148	649	–
upMaggr2	unknown protein	–	*Magnaporthe grisea*	XP_365988	353	–
upMaggr3	unknown protein	–	*Magnaporthe grisea*	XP_365989	600	–
(Black of Fig. 2)						
upArath	unknown protein	–	*Arabidopsis thaliana*	AAL15255	306	–
upBacag	unknown protein	–	*Bacillus agaradhaerens*	CAD38091	714	–
upBurps	unknown protein	–	*Burkholderia pseudomallei*	CAH37589	871	–
upCloac	unknown protein	–	*Clostridium acetobutylicum*	AAK80197	170	–

Table 1. (Continued).

Abbreviation	Specificity	EC number	Source	GenBank	Length	Glycoside hydrolase family
upCrypa	unknown protein	–	*Cryptosporidium parvum*	EAK89630	150	–
upDicdi	unknown protein	–	*Dictyostelium discoideum*	AAO51512	146	–
upDrome	unknown protein	–	*Drosophila melanogaster*	AAF46674	679	–
upGlovi	unknown protein	–	*Gloeobacter violaceus*	BAC91285	845	–
upHomsa	unknown protein	–	*Homo sapiens*	AAH27588	672	–
upChrvi	unknown protein	–	*Chromobacterium violaceum*	AAQ61151	874	–
upMusmuH	unknown protein	–	*Mus musculus* (head)	BAC31004	675	–
upMusmuL	unknown protein	–	*Mus musculus* (liver)	BAC34244	338	–
upMusmuT	unknown protein	–	*Mus musculus* (tymus)	BAC27063	128	–
upOrysa1	unknown protein	–	*Oryza sativa*	BAB63700	379	–
upOrysa2	unknown protein	–	*Oryza sativa*	AAU10756	373	–
upRatno	unknown protein	–	*Rattus norvegicus*	AAO84024	672	–
upXenla	unknown protein	–	*Xenopus laevis*	AAH73202	313	–
CBM21						
(Bright green of Fig. 2)						
amyLipko	α-amylase	3.2.1.1	*Lipomyces kononenkoae*	AAC49622	624	13
amyLipst	α-amylase	3.2.1.1	*Lipomyces starkeyi*	AAN75021	647	13
(Blue of Fig. 2)						
gmyArxad	glucoamylase	3.2.1.3	*Arxula adeninivorans*	CAA86997	624	15
gmyRhior	glucoamylase	3.2.1.3	*Rhizopus oryzae*	AAQ18643	604	15
gmyMucci	glucoamylase	3.2.1.3	*Mucor circinelloides*	AAN85206	609	15
(Pink of Fig. 2)						
pfHomsa	protein phosphatase	3.1.3.16	*Homo sapiens*	AAB94596	1122	–
pfRatno	protein phosphatase	3.1.3.16	*Rattus norvegicus*	CAA77083	284	–
pf_MusmuA	protein phosphatase	–	*Mus musculus* (adipocyte cells)	AAB49689	294	–
pf_MusmuH	protein phosphatase	–	*Mus musculus* (heart)	AAK31072	578	–
pf_MusmuL	protein phosphatase	–	*Mus musculus* (lungh)	AAH60261	284	–
pfrsGalga	prot. phosp. reg. sub.	–	*Gallus gallus*	AAC60216	288	–
pfrsHomsaB	prot. phosp. reg. sub.	–	*Homo sapiens* (brain)	AAH47502	299	–
pfrsOrycu	prot. phosp. reg. sub.	–	*Oryctolagus cuniculus*	AAA31462	1109	–
pfrsSacce1	prot. phosp. reg. sub.	–	*Saccharomyces cerevisiae*	CAA86906	538	–
pfrsSacce2	prot. phosp. reg. sub.	–	*Saccharomyces cerevisiae*	CAA45371	793	–
pfrs_Cloac	prot. phosp. reg. sub.	–	*Clostridium acetobutylicum*	AAK76874	247	–
pfrs_HomsaS	prot. phosp. reg. sub.	–	*Homo sapiens* (skin)	AAH43388	285	–
pfrs_HomsaM	prot. phosp. reg. sub.	–	*Homo sapiens* (muscle)	AAH12625	317	–
pfrs_Sacce1	prot. phosp. reg. sub.	–	*Saccharomyces cerevisiae*	AAB64590	548	–
pfrs_Sacce2	prot. phosp. reg. sub.	–	*Saccharomyces cerevisiae*	AAB67365	648	–
pfrs_Xentr	prot. phosp. reg. sub.	–	*Xenopus tropicalis*	AAH74693	223	–
(Black of Fig. 2)						
upAspni	unknown protein	–	*Aspergillus nidulans*	EAA64131	795	–
upCaeel1	unknown protein	–	*Caenorhabditis elegans*	AAF39789	318	–
upCaeel2	unknown protein	–	*Caenorhabditis elegans*	AAK82903	346	–
upCangl1	unknown protein	–	*Candida glabrata*	CAG59109	682	–
upCangl2	unknown protein	–	*Candida glabrata*	CAG59903	915	–
upCangl3	unknown protein	–	*Candida glabrata*	CAG60779	543	–
upCangl4	unknown protein	–	*Candida glabrata*	CAG61779	827	–
upDanre1	unknown protein	–	*Danio rerio*	AAH44421	293	–
upDanre2	unknown protein	–	*Danio rerio*	AAH67184	253	–
upDanre3	unknown protein	–	*Danio rerio*	AAH75881	311	–
upDanreW	unknown protein	–	*Danio rerio* wild-type	AAH60926	317	–
upDebha1	unknown protein	–	*Debaryomyces hansenii*	CAG87286	628	–
upDebha2	unknown protein	–	*Debaryomyces hansenii*	CAG89742	509	–
upDrome1	unknown protein	–	*Drosophila melanogaster*	AAF49732	330	–
upDrome2	unknown protein	–	*Drosophila melanogaster*	AAF49172	172	–

Table 1. (Continued).

Abbreviation	Specificity	EC number	Source	GenBank	Length	Glycoside hydrolase family
upErego1	unknown protein	–	*Eremothecium gossypii*	AAS51837	354	–
upErego2	unknown protein	–	*Eremothecium gossypii*	AAS54765	679	–
upHomsaR	unknown protein	–	*Homo sapiens* (retina)	CAD97641	317	–
upHomsaS	unknown protein	–	*Homo sapiens* (spleen)	BAB15779	349	–
upKlula1	unknown protein	–	*Kluyveromyces lactis*	CAH00570	748	–
upKlula2	unknown protein	–	*Kluyveromyces lactis*	CAG99013	498	–
upMaggr	unknown protein	–	*Magnaporthe grisea*	XP_367749	924	–
upMusmu	unknown protein	–	*Mus musculus*	AAF66954	735	–
upNeucr	unknown protein	–	*Neurospora crassa*	XP_330896	864	–
upXenla1	unknown protein	–	*Xenopus laevis*	AAH72880	271	–
upXenla2	unknown protein	–	*Xenopus laevis*	AAH68825	223	–
upXenla3	unknown protein	–	*Xenopus laevis*	AAH77483	299	–
upXenla4	unknown protein	–	*Xenopus laevis*	AAH73501	313	–
upYarli	unknown protein	–	*Yarrowia lipolytica*	CAG82944	1129	–

however, that the fold recognition method 3D-PSSM [61] identified the CBM20 module of *Bacillus stearothermohilus* maltogenic α-amylase [62] as a top hit for CBM21 SBDs from both *R. oryzae* glucoamylase [49] and *Lipomyces kononenkoae* α-amylase [63]. In addition, secondary structure prediction for these two SBDs from CBM21 indicates that β-strands would be expected to occur in positions equivalent to known β-strand locations in CBM20 domains, when the amino acid sequences are aligned as in Fig. 2. These findings, together with the secondary structure prediction of the glycogen-targeting subunit of protein phosphatases [50], strongly support the idea that the three-dimensional structures of CBM20 and 21 modules are similar and suggest that the two CBM families can be grouped into a CBM clan.

Compared to CBM20, analysis of CBM21 sequences received much less attention [24,50,64]. Based on the present alignment, it is clear that some of the CBM20 consensus residues, Gly628, Trp643, Trp689 and Asn694 (*B. circulans* CGTase numbering including the signal peptide) have possible equivalents in the CBM21motif (Fig. 2). Concerning Trp663 (i.e., Trp636 without the signal peptide), which possesses a structural role in CBM20 instead of a binding role [65], this residue is evidently present in all amylolytic CBM21 SBDs (from recognized α-amylases and glucoamylases). The remaining CBM21 sequences contain a phenylalanine in that position (Fig. 2), with the exception of the regulatory subunit of protein phosphatase from *Clostridium acetobutylicum* (that moreover contains the lysine equivalent to the CBM20 consensual Lys678, i.e., Lys651 without the signal peptide). Interestingly, the two tryptophans (corresponding with the two functional CBM20 Trp residues) are better conserved in the nonamylolytic CBM21 motifs than in CBM21 SBDs from α-amylases and glucoamylases (Fig. 2).

Evolutionary analysis

The evolutionary relationships between the numerous CBM20 and CBM21 sequences (Table 1) are apparent in Fig. 4. The two families clearly retain some independence, thus CBM20 members do not occur in the CBM21 part of the tree and vice versa. In the past, by far the most attention was paid to the evolution of

Fig. 2. Alignment of SBD sequences from CBM20 and CBM21 families. For an explanation of the colour code for enzymes and the abbreviations used for the sources, see Table 1. Only the segments around the important residues (known as consensus [23]; blue and yellow highlighting) plus the one at the beginning of the SBD modules are shown. In the CBM20 module, the tryptophans and tyrosines involved in binding sites 1 and 2, respectively, are signified by yellow [41,42]. The conserved phenylalanine in CBM20 and invariant lysine in CBM21 are shown in black inversion. The aspartate and two phenylalanines (DxFxF) in CBM21, characteristic of nonamylolytic enzymes, are highlighted in gray. The numbers preceding the first segment and succeeding the last segment represent the position in the amino acid sequence. Residues deleted between the two adjacent segments are indicated by superscript numbers. The sequences are numbered from the N-terminus including the signal peptides (e.g. for CGTase from *Bacillus circulans* strain 251, there is a known 27-residue long signal peptide). The two extra lines under each CBM family, 90% cons and 80% cons, are associated with 90% and 80% consensus, respectively. Special symbols are used for aromatic (▲), acidic (△), hydrophobic (●), and hydrophilic (○) residues.

CBM-20

```
amyAspka     540 PITFEEL  2 TTYGE 1 VYLSGSISQ-LGEWHT  8 DDYTSSNPEWSV  9 FEYKFIK  8 WESDP--N  619
amyAspnd     523 TVVFQER  2 TAYGE 1 VFLAGSISQ-LGNWDT  8 AQYTATDPLWTV  9 FEFKFLK  8 WESNP--N  602
amyBacsp     517 NVTFTVN  3 TTSGQ 1 VYVVANIPE-LGNWNT  9 -----SYPTWKA  9 IEFKFIK  8 WESTS--N  593
amyCrysp     530 TVVFDVY  2 TQYGQ 1 VVIAGNIPQ-LGNWSP  9 -QYTASSPKWTG 10 FQWKPIV  7 WYPGN--N  609
amyStrgr     472 SASFHVN  2 TAWGE 1 IYVTGDQAA-LGNWDP  7 ---PAAYPVWKL  9 FQYKYLR  8 WESGA--N  547
amyStrlm     472 SASFHVN  2 TAWGE 1 IYVTGDQAA-LGNWDP  7 ---PAAYPVWKL  9 FQYKYLR  8 WESGA--N  547
amyStrli1    480 RRVPSAV  3 TSWGQ 1 IYVTGNRPE-LGNWNP  7 ---PAAYPVWKR  9 FEYKYLR  8 WESGA--N  556
amyStrli2    480 GVSFAVD  2 TSWGQ 1 IYVTGNRPG-LGHWDP  7 ---PAAYPVWKR  9 FEYKSLR  8 WESGA--N  555
amyStrvi     475 GASFNVT  2 TVVGQ 1 IYVTGNRAE-LGNWAP  7 ---PATYPVWKL  9 FEYKYIR  8 WESGA--N  550
amyThncu     507 TARFHAT  2 TWYGQ 1 VAVVGSIPE-LGSWQP  8 --DSGTYPVWSG  9 FEYKYVK  6 WSGSR--A  582
amy_Aspaw    536 PITLEEL  2 TTYGE 1 IYLSGSISQ-LGEWDT  8 DDYTSSNPEWYV  9 FEYKFIK  8 WESDP--N  615
atrActsp     624 PVQFTVQ  4 TAPGE 1 LYLTGDVAE-LGHWST  8 LLRVPNESRGVL  9 VEFKFVK  8 WEGGA--N  705
cgtBacag     582 TVRFIID  3 TKLGE 1 VFLVGNVHE-LGNWDP  9 -QIVYQYPTWYY  9 LEFKFIK  8 WQSGA--N  662
cgtBacbr     595 SVRFAVN  3 TNSGT 1 VYIVGNVSE-LGNWDP  9 -QVMYKYPTWYY  9 LEYKYIK  8 WQSGN--N  675
cgtBacci2    615 SVRFVVN  3 TALGQ 1 VYLTGSVSE-LGNWDP  9 -QVVYQYPNWYY  9 IEFKFLK  7 WEGGS--N  694
cgtBacci8    620 TVRFVVN  3 TTLGQ 1 LYLTGNVAE-LGNWST 10 -QVIHQYPTWYY  9 LEFKFFK  7 WESGS--N  700
cgtBacciA    615 TVRFVIN  3 TALGQ 1 VFLTGNVSE-LGNWDP  9 -QVVYQYPTWYY  9 IEFKFLK  7 WEGGA--N  694
cgtBaccl     603 SVRFVVD  3 TNYGE 1 VYLVGNVPE-LGNWNP  9 -QVVYSYPTWYY  9 LEFKFII  8 WESGG--N  683
cgtBacli     620 SVRFVIN  3 TALGE 1 IYLTGNVSE-LGNWTT 10 -QVIHAYPTWYY  9 LEFKFFK  7 WEGGS--N  700
cgtBacma1    616 TVRFLVN  3 TNYGT 1 VYLVGNAAE-LGSWDP  9 -QVIAKYPSWYY  9 LDFKFIK  7 WEGGG--N  695
cgtBacma2    615 TVRFKVN  3 TALGQ 1 VYLTGNVAE-LGNWTA  9 -QVEASYPTWYF  9 LQFKFIK  7 WEGGN--N  694
cgtBacoh     606 SIRFAVN  3 TSLGT 1 LYMVGNVNE-LGNWDP  9 -QVMYQYPTWYY  9 LEYKFIK  8 WESGN--N  686
cgtBacsp0    615 TVRFVIN  3 TALGQ 1 VFLTGNVSE-LGNWDP  9 -QVVYQYPTWYY  9 IEFKFLK  7 WEGGA--N  694
cgtBacsp1    606 SVRFGVN  3 TSPGT 1 LYIVGNVNE-LGNWDA  9 -QVMYQYPTWYY  9 LEYKYIK  8 WQSGN--N  686
cgtBacsp7    615 SVRFVVN  3 TALGQ 1 VYLAGSVSE-LGNWDP  9 -QVIYQYPTWYY  9 IEFKFLK  7 WEGGS--N  694
cgtBacsp3    614 TVRFVIN  3 TALGQ 1 VFLTGNVSE-LGNWDP  9 -QVVYQYPTWYY  9 IEFKFLK  7 WEGGA--N  693
cgtBacsp63   620 TVRFVIN  3 TTLGQ 1 IYLTGNVAE-LGNWST 10 -QVIHQYPTWYY  9 LEFKFFK  7 WEGGS--N  700
cgtBacsp6    606 SIRFAVN  3 TSLGT 1 LYIVGNVNE-LGNWDP  9 -QVMYQYPTWYY  9 LEYKFIK  8 WESGN--N  686
cgtBacspB    615 SVRFVVN  3 TALGQ 1 LYLTGNVSE-LGNWDP  9 -QVVYQYPNWYY  9 IEFKFLK  7 WEGGS--N  694
cgtBacspD    600 SVRFVVN  3 TSVGE 1 LYVVGDVPE-LGSWDP  9 -QVLYSYPTWYY  9 IEYKYIM  8 WESGN--N  680
cgtBacspE    679 SVRFGVN  3 TSPGT 1 LYIVGNVNE-LGNWDA  9 -QVMYQYPTWYY  9 LEYKYIK  8 WQSGN--N  759
cgtBacspK    628 SVRFGVN  3 TSPGT 1 LYIVGNVNE-LGNWDA  9 -QVMYQYPTWYY  9 LEYKYIK  8 WQSGN--N  708
cgtBacst     612 SVRFVVN  3 TNLGQ 1 IYIVGNVYE-LGNWDT  9 -QVVYSYPTWYI  9 IEFKFIK  8 WESGS--N  692
cgtGeost     612 SVRFVVN  3 TNWGE 1 IYLVGNVHE-LGNWNT  9 --VIYSYPTWYV  9 IEFKFIK  8 WESGS--N  692
cgtKlepn     561 SINFTCN  3 TISGQ 1 VYIIGNIPQ-LGGWDL  7 --PTQ-YPQWSA  9 VEWKCVK 11 WQSGA--N  640
cgtThmth     612 CVRFVVN  3 TVYGE 1 VYLTGNVAE-LGNWDT  9 -QVVYQYPTWYY  9 IQFKFIK  7 WEGGS--N  691
cgtThcsp     636 PAIFEVR  8 TQVGE 1 LWLTGSVPE-LSYWSP  7 PMLCPGWPDWFV  9 IEFKFLK  8 WEVGS--N  720
cgt_Bacsp5   614 TVRFVIN  3 TALGQ 1 VFLTGNVSE-LGNWDP  9 -QVVYQYPTWYY  9 IEFKFLK  7 WEGGA--N  693
cgt_Glovi    544 IVRIQVN  3 TQPGE 1 VAVIGDCPE-LGDWDL 10 ------DNTWFG 11 VAYKYVI  7 INENRT-S  622
cgt_Nossp7   541 IVRVQLN  3 TQPGE 1 IVVVGDCPE-LGNWDI 10 ------SNTWFA 11 ISYKYAM  7 LRENIL-N  619
cgt_Nossp9   541 IVRAQLN  3 TQPGE 1 IVVIGDCPE-LGNWDI 10 ------TNTWFA 11 IAYKYAL  7 LRENLV-N  619
cgt_Stcpy    613 PVRLLIN  3 TVPGE 1 LYLMGDVFE-MGANDA 10 TQTIAKYPNWFF  9 IAVKLVK  8 WTSPE--T  695
m5hPsespK    516 SLTFNET  2 TVWGQ 1 LFVVGNVGA-LGNWAP  8 ISGSGSTGQWRA  9 VQYKYVK  8 WESGG--N  595
m4hPsesa     456 NVNFRCD  3 TQMGD 1 VYAVGNVSQ-LGNWSP  7 --DTSSYPTWKG  9 VEWKCLI 11 WQSGG--N  536
m4hPsest     452 SVSFRCD  3 TQMGD 1 VYAVGNVSQ-LGNWSP  7 --DTSGYPTWKG  9 EEWKCLI 11 WQGGA--N  532
maaBacst     614 SVVFTVK  3 TNLGD 1 IYLTGNIPE-LGNWST 11 PLLAPNYPDWFY  9 IQFKFFI  8 WENGS--N  698
apuBacst    1341 QVTFKVR  2 SYTPL 2 RITIPNSLN--G-WNT  7 -G-GAVTSDWEF  9 IIYKYVK 26 YGAIGT-D 1435
apuBacspX   1337 QVTFKVK  2 SYTPL 2 RITIPNSIN--G-WNT  7 -G-GAVTPDWEF  9 ITYKYVK 26 YGAIGT-E 1431
apuTheth    1253 KVIFNVT  2 DYTPD 1 VNLAGTFPN--ATWDP  7 -I---DNNTYSI  9 IEYKYAR 10 YGNEFASN 1330
apuTheet    1256 KVIFNVT  2 DYTPD 2 ANIAGNFHD--AFWNP  8 -----GPNTYSI  9 LEYKYAR 10 YGEEIA-N 1333
apuThetc    1261 KVTFNVT  2 DYTPD 2 VNIAGNFPD--AFWNP  8 -----GSNTYSI  9 IEYKYAR 10 YGNEID-N 1338
bmyBacce     451 MQTIVVK  3 TTIGD 1 VYITGNRAE-LGSWDT 10 ----SHSNDWRG  9 IEFKAFI  9 WQTIQ--Q  530
bmyBacme     451 AQTVVVK  3 TALGE 1 VYIVGDRAE-LGQWDT 10 ----SSTADWRG  9 VQFKAIV  9 WQPSQ--Q  530
bmyCloth     455 PVTFTIN  3 TYYGQ 1 VYIVGSTSD-LGNWNT 10 -----NYPTWTI  9 IQFKAVK  8 WEGGS--N  532
gmyAspaw     539 AVTFDLT  2 TTYGE 1 IYLVGSISQ-LGDWET  8 DKYTSSNPLWYV  9 FEYKFIR  8 WESDP--N  618
gmyAspfi     540 AVTFDLT  2 TTYGE 1 IYLVGSISQ-LGDWET  8 DKYTSSDPLWYV  9 FEYKFIR  8 WESDP--N  619
gmyAspka     539 AVTFDLT  2 TTYGE 1 IYLVGSISQ-LGDWET  8 DKYTSSNPLWYV  9 FEYKFIR  8 WESDP--N  618
gmyAspni     540 AVTFDLT  2 TTYGE 1 IYLVGSISQ-LGDWET  8 DKYTSSDPLWYV  9 FEYKFIR  8 WESDP--N  619
gmyAspor     513 SVTFAVK  2 TVYGE 1 IKIVGSISQ-LGSWNP  8 DSYTTDNPLWTG  9 FEYKFIR  7 WESDP--N  591
gmyAspsh     539 AVTFDLT  2 TTYGE 1 IYLVGSISQ-LGDWET  8 DKYTSSNPPWYV  9 FEYKFIR  8 WESDP--N  618
gmyAspte     616 AVTFDEV  2 TTYGE 1 VYVVGSISQ-LGSWDT  8 SKYTSSNNLWYV  9 FQYKFIR  8 WESDP--N  695
gmyCorro     484 EVTFDVY  2 TVYGQ 1 IYITGDVSE-LGNWTP  7 ---SANYPTWSA  9 IQYKYVN  7 WEDAIS-N  559
gmyHorre     508 SITFNIN  2 TYYGE 1 LYVIGNSSD-LGAWNI  8 SAYTQDRPLWSA  9 ISYQYVR  7 YIYETV-N  587
gmyHumgr     516 YVTFNER  2 TAWGE 1 IKVVGNVPA-LGNWDT  8 SGYKSNDPLWSI 10 VQYKYIK  8 WESDP--N  596
gmyLened     478 SVTFNVD  2 TLEGQ 1 VYLTGAVDA-LEDWST  7 ---SANYPTWSV  9 VQYKYIK  8 WESDP--N  553
gmyNeucr     527 LVTFNEK  2 TSYGQ 1 VKVVGSIAA-LGNWAP  8 KQYSSSNPLWST  9 FKYKYVV  8 WENDP--D  606
gmyTalem     491 AVTFDEI  2 TSYGE 1 IYLAGSIPE-LGNWST  8 DAYTNSNPLWYV  9 FEYKFFK  8 WEDDP--N  570
gmy_Aspaw    539 AVTFDLT  2 TTYGE 1 IYLVGSISQ-LGDWDT  8 DKYTSSNPLWYV  9 FEYKFIR  8 WESDP--N  618
gmy_AspniT   539 AVTFDLT  2 TTYGE 1 IYLVGSISQ-LGDWET  8 DKYTSSDPLWYV  9 FEYKFIR  8 WESDP--N  618
gmy_Neucr    305 AVTFNHL  2 TSYGE 1 IKIVGSISQ-LGSWSA  8 SQYTTSNPLWTA  9 FEYKFVK  9 WESDP--N  385
6agtArtgl    866 WATFSCE  3 TTFGQ 1 VYVVGNVPQ-LGNWSP  7 --PSA-YPTWTG 10 VEWKCIK 12 WEPGG--N  947
4agtBacfr    149 -------  0 ----- 1 LAICGNQKA-LGNWDP  8 ----ANFPEWQA 10 LEYKFVL 10 WENNP--N  212
4agtSoltu      9 KVSFRIP  2 TQWGQ 1 LLICGSDRL-LGSWNV  8 -SHQGEVLVWSI  7 SEYSYYV  9 WEVGK--K   89
4agt_Arath   164 VVQFKIC  3 IGEGT 1 VYVLGTPEK-LGNWKV  7 ---YVDDSIWEA 10 IKYRYCK  8 FESGG--N  241
4agt_Orysa    19 TLVFKLP  2 TQWGQ 1 LLIAGSEPA-LGSWNV  8 -VHQGNELIWSG  9 CQYNYYV  9 SESGE--K   98
agwdArath     75 RLNVRLD  2 VNFGD 1 VAMFGSAKE-IGSWKK  8 -------NGWVC  9 LECKFVI  8 WESGD--N  147
genHomsa     265 SVRFQVH  3 STDVQ 1 IAVTGDHEC-LGRWNT  7 -----KDGFWSH  9 VEWKFVL  8 WEECS--N  339
lafGalga       2 LFRFGVV  5 AEGGG 1 LLVAGSRPE-LGEWDP 13 ALAAQEPVLWLG 12 FWYKFLR  7 WEGNG--P   91
lafHomsa       2 RFRFGVV  6 GAR-P 1 LLVVGSRPE-LGRWEP 16 ALALQEPGLWLG 21 FWYKFLK  8 WEGNG--P  104
depChlpr      23 KVQFRLP  2 VSFGQ 1 ISIVTS----RSGWEP  7 ---WSEGDEWKV  9 LEYKYVV  8 WQTGS--N   95
```

```
upAspnd1    285 PVTFNAL  2 TTYGE  1 VYLAGSISQ-LGSWST  8 SKYSSSSPLWTV  9 FEYKYIK  8 WESGP--N 364
upAspnd2    551 AVTFNVI  2 TTYGE  1 VYIVGSISQ-LGNWDT  8 SKNTSSNNLWYV  9 FEYKYIR  8 WESDP--N 630
upAspnd3    520 SVTFELT  2 TVWGE  1 IRLVGSSGE-LGYWVS  8 DRYSSSQPIWWA  9 VEYKYIR  8 LRYNN--T 599
upMaggr1    549 AVTFNVL  2 TTPGD  1 IKIVGDIED-LGKWNP  8 NDYTASRPLWKK  9 VQYKFIN  8 WEADP--N 628
upMaggr2    254 AVTFRSK  2 TSVGQ  1 VKIAGSIAQ-LGGWDA  8 SQYTSSNPLWTT  9 FEYKFIR  8 YESGA--N 333
upMaggr3    502 AVSFTVR  2 TAPGD  1 IKMVGNTAQ-LGSWDA  8 SGYNSTNMAWSI  9 VQYKFVK  8 WESDP--N 581
upArath      88 RVRFQLR  2 CVFGE  1 FFIVGDDPVFGGLWDP  7 ---WSDGNVWTV  9 VEFKLLL  8 WQPGP--N 164
upBacag     616 TVRFIID  3 TKMGE  1 IFLVGNVHE-LGNWDP  9 -QVVYQYPTWYY  9 LEFKFIK  8 WQSGA--N 696
upBurps     776 PVAVNVN  2 TQLGQ  1 MYVTGNVAA-LGNWNT  7 ---PASYPVWRN  9 IQYKYYR  8 WENRSG-N 852
upCloac      72 DVTFILN  3 TSIGE  1 IFISGNIKE-LGNWTI  7 -TDESIYPTWKT  9 VEFKFLL 11 WENSG--N 153
upCrypa      54 IVKFGVK  2 TKFGQ  1 LKVVGNIAE-LGNWNV  7 ---WTEGSFWTA 12 IEYKYVL  9 WEPGK--N 133
upDicdi      14 NVTFTIV  2 TQPGE  1 LFITGSMNQ-LGEWDT  9 -----IGNLWDA  9 IQYKYFV  9 WESIE--N  90
upDrome      53 VFNFTLT  6 LASFE  1 PALVGNLPV-LGAWQA  8 ---TAILNVWSA  9 VEYRYFA 13 WESHV--Q 138
upGlovi      88 VYRFQLI  2 TQIGE  1 IGLVGSAPE-LGRWDV  8 --SAKLYPLWRT 17 VEYKYIR  8 WEAIGP-D 174
upHomsa       5 QVAFEIR  2 LLPGE  1 FAICGSCDA-LGNWNP  8 ENDTGESMLWKA  9 VQYRYFK 19 WETHL--Q  95
upChrvi     779 AVAINAS  2 TQWGQ  1 VYVTGNARA-LGNWNT  9 -----AYPSWKN  9 IAYKYYR  8 WENLAG-N 855
upMusmuH      1 ----GPA  2 GAR-Q  2 LLLAGSRPE-LGRWEP 16 ALALQEPGLWLA 20 FWYKFLQ  8 WEGTA--L  94
upMusmuL    245 SIQFQVH  3 NTDVQ  1 IAVTGDHES-LGRWNT  7 -----KDGLWSH  9 VEWKFVL  8 WEECS--N 319
upMusmuT      5 QVTFEIR  2 LLPGE  1 FAICGSCDA-LGNWNP  8 ENETGDSVLWKA  9 VKYRYFR 19 WETHL--Q  95
upOrysa1    123 HVKFVLQ  2 CAFGQ  1 FLVVGDVAA-LGLWNP  7 ---WSEDHVWTV  9 IEFKFLL  8 WQHGR--N 198
upOrysa2    101 RVRFVLK  2 CTFGQ  1 FHLVGDDPA-LGLWDP  7 ---WSEGHDWTV  9 IEYKFVL  8 WQNGR--N 176
upRatno       5 QVTFEIR  2 LLPGE  1 FAMCGNCDA-LGNWSP  7 ESETGES-VWKA  9 VKYRYFR 19 WETHL--Q  93
upXenla       2 LFRFGVV  5 SDNE-  0 LLVLGSRKE-MGSWDP 15 ----TEPSFWVG 11 FWFKFIK  7 WEGNG--P  86

90% cons                                G     LG W               W         K       ▲      ○
80% cons           F          G         G     LG W               W         K       W      N
```

CBM-21

```
amyLipko     48 VQLASYE  5 -LSAS  9 KIVTLYYLTS--SGTT  8 PVWSNNWELWTL  5 --GAVEI  5 VDSDTSVT 129
amyLipst     48 VQLASHE  5 -LSAS  9 KIVTLYYLTS--SGTT  8 SSLSNNWELWSL  5 --DAVEI  5 VDSDASAT 129
gmyArxad     34 VALSSYS  5 -LSAS  9 KLVTLYWTNA--DNKS 13 ASDDQSWELWSL 12 LNITYVA  4 KTNSQQLN 128
gmyRhior      9 VQLDSYN  5 -FSGK  9 KKVTVIYADGSDNWNN 13 --SGSNYEYWTF  9 FYIKYEV  5 YDNNNSAN 101
gmyMucci     28 PTTEAVK  5 -LAGQ  9 KTVTVIYSDASDNWNN 13 --AGTNYEYWTF  9 FYVKYVV  5 YDNNNSGN 125
pfHomsa     129 ILESTES  5 SIKGI  9 KLVYVRMS--LDDWQT 12 --CDGETDQFSF 16 FCIRYET  5 WSNNNGTN 226
pfRatno     132 VCLENCV  5 -IAGT  9 KVVKISMT--FDTWKS 11 TYAGSDRDTFSF 15 FAVCYEC  5 WDSNKGKN 228
pf_MusmuA   134 VCLENCS  5 -VTGT  9 KKVQVRIT--FDTWKT 11 VYSSSDSDTFSF 15 FCISYHA  5 WDNNEGQN 230
pf_MusmuH   131 VLESAEH  4 SMKGI  9 KLVYVRMS--LDDWQT 12 --CDGETDQFSF 16 FCIRYET  5 WSNNNGTN 227
pf_MusmuL   165 VCLENCV  5 -IAGT  9 KVVKIRMT--FDTWKS 11 TYAGSDRDTFSF 15 FAVCYEC  5 WDSNKGKN 261
pfrsGalga   140 VCLESCL  5 -LSGT  9 KKVLVRIT--FDGWKS 11 TYGSADMDTFSF 15 FCISFRC  5 WDNNQGKN 236
pfrsHomsaB  177 VCLERVT  5 -ISGT  9 KQVAVRYT--FSGWRS 14 EGT---EDVFTF 16 FAVRYQV  5 WDNNDHRD 274
pfrsOrycu   131 MLESTEY  5 SMKGI  9 KLVYVRMS--LDDWQT 12 --CDGETDQFSF 16 FCIRYET  5 WSNNNGTN 228
pfrsSacce1  392 VFLQEIT  8 VIIGK  9 KKIIVRYT--WDAWRT 14 ILPGSNMDIFKF 19 FCIQYLT 10 WDNNDSAN 504
pfrsSacce2  244 VKLHSLT  8 -ITGL  9 KYLEIKFT--FNSWRD 11 --INSNVDEFKF 31 LCCRYDV  5 YDNNNGKN 357
pfrs_Cloac   44 QRITDDE  5 -VEGY  9 KNVYVHYS--LDGGKN 12 --PSDNYEVWKF 14 YCFKYEV  5 WDNNNGKN 138
pfrs_HomsaS 133 VCLENCV  5 -IAGT  9 KTVKIRMT--FDTWKS 11 TYAGSDRDTFSF 15 FAVYYEC  5 WDSNRGKN 229
pfrs_HomsaM 157 VCLENCS  5 -VTGT  9 KKVQIRIT--FDSWKN 11 VYGGTDSDTFSF 15 FCISYHA  5 WDNNDGQN 253
pfrs_Sacce1 424 CNGVAKG  7 LIAGR  9 KRVVVRYT--WDSWRT 14 ILPGTNMDIFHF 15 FCIHYST  9 WDNNNGNN 530
pfrs_Sacce2 209 VSYEDIC  8 -IWGL  9 KKIEIKFT--LNNWAD 11 --VTPHVDEFKF 33 FCCRYDV  8 YDNNDYKN 327
pfrs_Xentr  112 VCLEQCA  5 -VAGT  9 KRVTLRVS--YDGWCN 15 ----GDTDSFSF 12 FCICYWC  5 WDNNDGKN 205
upAspnd     440 LERLFLS  5 -LVGQ  9 KHVAARFT--FDNWRT 14 KQLHDGYDRFMF 16 VCIRYNV  5 WDNNETRN 540
upCaeel1    135 VCVAALR  5 -IVGQ  9 KVVVVRYT--IDGWAT 14 TED---IDAFNF 14 FCVQYQV  5 WDNNGGDN 230
upCaeel2    247 VSLENVI  6 KVMGT  9 KSVFVRYT--MNGWIS 11 TSK--IQDTFKF 15 FCICFKA  5 WDSNSGTN 343
upCangl1    177 GSNIKLH  8 -LKGL  9 KFIEVKFS--FNNWKD 11 --ITDHIDEFRF 35 LCCRYDV  5 YDNNNYKN 294
upCangl2    286 VSLAQDS  5 -IVGK  9 KFIEVKYT--FNNWND 11 --ISAEIDEFEF 29 LCCRYDV  5 YDNNNYRN 394
upCangl3    419 SGVDIGT  6 -LSGR  9 KRVLIRYT--WDRWRN 19 ----AAMDVFHF 15 FCIQYTT  9 WDNNCGKN 524
upCangl4    710 IGFSTGR  7 -ITGT  9 KKVSIRYT--WDHWRS 18 ----SQMDIFRF 21 FCIQYTT  9 WDNNNGKN 821
upDanre1    137 VCLEHCM  5 -IMGT  9 KSVKLRIT--FNTWKN 11 TYTGSNRDTFSF 15 FAICYEV  5 WDNNQGKN 233
upDanre2    148 VLLESCN  5 -VLGT  9 KAVHVRIT--FDSWKT 11 CYGEPGTDVFEF 15 FCVSYLP  7 WDNNNGKN 246
upDanre3    147 VCLENCT  5 -LTGT  9 KSVHVRIT--FDSWKS 15 ----QDTDTFSF 15 FCISFRT  5 WDNNDGRN 243
upDanreW    158 VCLENCI  5 -VTGT  9 KVVHVRIT--FDSWKS 11 VYGCEDVDTFSF 15 FCLSYKT  5 WDNNDGKN 254
upDebha1    273 VFLERIF  7 -LLGH  9 KFITVRYT--LDNWCT 19 -----NYDRFIF 30 LCIKYNT  5 WDNNESRN 389
upDebha2    196 YLQSIKL  6 -LVGL  9 KRLSIKLT--FNAWRS 16 -----NFDQFKF 14 FVIKYEV  5 WDNNNLKN 292
upDrome1    230 VSLENVI  6 IVVGT  9 KEIIVRVT--WDDWKS 15 TCAHVVFDTFSF 11 FCICYRT  5 WDNNDGKN 329
upDrome2    531 VSLENAA  7 TISGS  9 KSVHIRYS--LDGWRS 12 --CDGFSDIFTF 14 FAVRFQC  5 WDNNYGAN 628
upErego1    244 VFLQDLS  5 VMTGR  9 KSVIVRYT--WDNWAH 14 VLPGKDMDLFEF 15 FCIRYQV  9 WDNNHGNN 348
upErego2    273 IRLNKVS  6 -IKGS  9 KFIEVKFS--FDEWKN 11 --VTSKVDEFQF 33 LCCRYDV  5 YDNNNYEN 386
upHomsaR    157 VCLENCS  5 -VTGT  9 KKVQIRIT--FDSWKN 11 VYGGTDSDTFSF 15 FCISYHA  5 WDNNDGQN 253
upHomsaS    232 ICLERAE  5 -VAGS  9 KRVSVRWS--ADGWRS 14 PPR---ADRFAF 12 FALRYRV  5 WDNNGGRD 325
upKLula1    458 KLHECNS  7 -LTGL  9 KFIEIKFS--FNGWKD 16 -------DEFQF 29 LCCRYDV  5 YDNNNYEN 568
upKlula2    382 ICHLQDL  8 -LIGN  9 KRVIVRYT--LDSWKS 21 ----IDIDVFQF 14 MCILYQT  9 WDNNSGQN 490
upMaggr     456 LERVWLS  5 -LVGS  9 KHVTCRFT--FDYWKT 14 KESDVGHDRFNF 16 FCVRYNV  5 WDSNGGAN 556
upMusmu      84 PGGSGVW  7 VVRGL  9 KAVHVRAS--HDGWAT 43 PDDGGCTDRFAF 15 FVVRYET  5 WANNHGRN 215
upNeucr     361 LERVWLS  5 -LIGS  9 KSVTCRFT--LDYWKT 14 SESPLGQDRFNF 16 FCIRYNV  5 WDNNNGMN 461
upXenla1    127 VCLENCM  5 -LVGT  9 KCVKIRIT--FDSWQT 15 ----SDKDTFSF 15 FAVCFDC  5 WDSNKGLN 223
upXenla2    112 VCLEQCA  5 -VAGT  9 KRVTLRVS--YNGWRN 15 ----GDTDSFSF 12 FCFCYWC  5 WDNHDGKN 205
upXenla3    138 VCLENCS  5 -VAGT  9 KSVKIRIT--FNTWKS 15 ----TDSDTFSF 15 FCISYES  5 WDNNDGQN 234
upXenla4    152 VCLENCS  5 -VAGT  9 KSVKIRIT--FNTWKS 15 ----TDIDTFSF 15 FCISYES  5 WDNNDGQN 248
upYarli     664 VYLENLF  6 NLVGH  9 KQVNVRYS--LDYWQT 19 -----GYDRFTF 18 LCVRYTA  5 WDNNGFQN 768

90% cons                      G     K            W             △ ▲ F               ▲      N
80% cons                      G     K         ●  W             D F F               ▲      N
```

Fig. 2. (Continued).

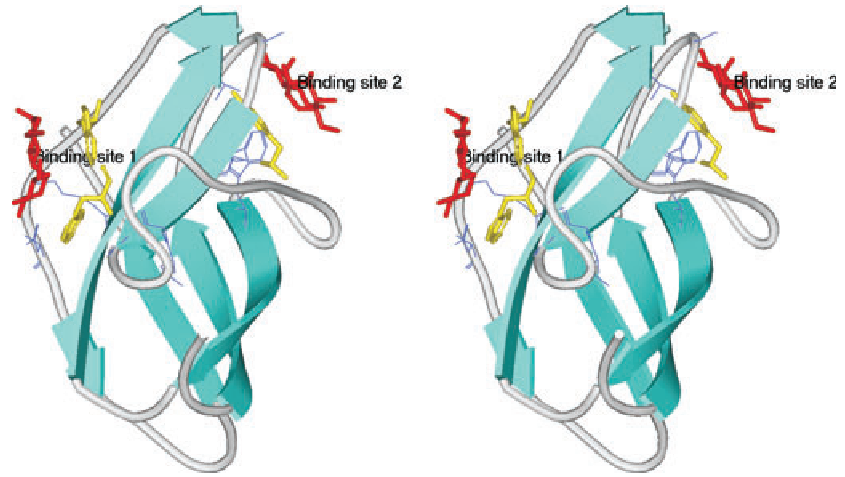

Fig. 3. The three-dimensional ribbon diagram of CBM20 module. The X-ray structure of SBD from *Bacillus circulans* strain 251 CGTase (PDB code: 1CDG [33]). The side chains of the aromatic residues involved in the starch-binding sites 1 (tryptophans) and 2 (tyrosines) are displayed in yellow in both SBDs. The two maltoses are shown in red. The nine further residues from the consensus SBD signature [23] are also displayed for comparison (in thin blue lines). Figure in stereo was prepared using the program WEBLABVIEWERLITE 4.0 (Accelrys Ltd, Cambridge, UK; http://www.accelrys.com/).

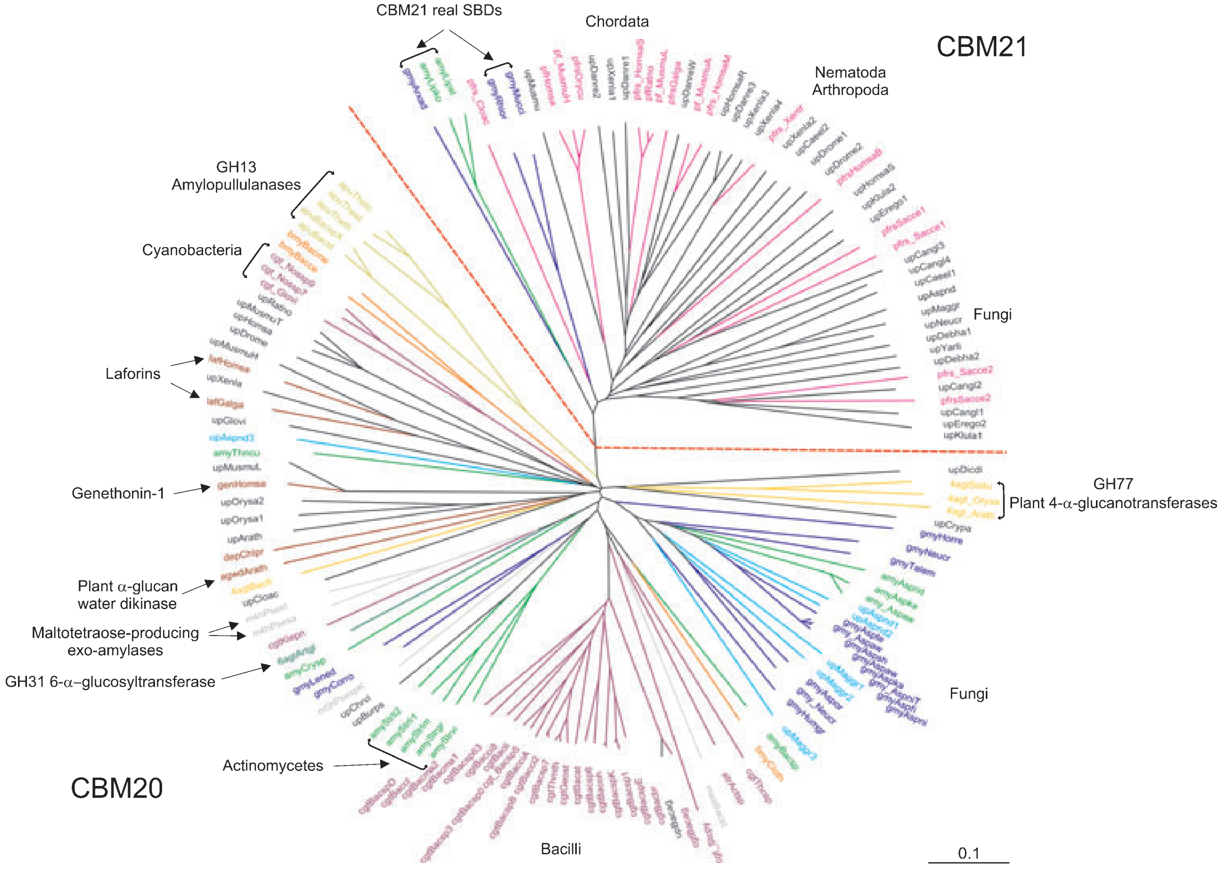

Fig. 4. Evolutionary tree of SBDs from CBM20 and CBM21. For an explanation of the colour code for enzymes and the abbreviations used for the sources, see Table 1. A red dashed line separates the CBM20 family from the CBM21. The tree is based on the alignment of complete SBD sequences including gaps.

CBM20 [24,25], and both families are studied together here for the first time.

The CBM21 part of the tree (Fig. 4) appears more compact than that of CBM20 perhaps simply due to the smaller number of CBM21 sequences. It may not be surprising that the known CBM21 SBDs from α-amylases and glucoamylases are located in two adjacent clusters positioned most closely to the borderline between the families (gmyArxad, amyLipko, amyLipst, gmyRhior, and gmyMucci). In other words these real SBD CBM21 modules are most closely related to the CBM20 family. Of the remaining nonamylolytic CBM21 sequences, only the module of the regulatory subunit of protein phosphatase from *C. acetobutylicum* (pfrsCloac) was found located clearly among the amylolytic SBDs, reflecting the sequence features discussed above. The rest of the remaining sequences form a large, more or less undifferentiated cluster that gives the possibility of identifying several related subgroups, such as Chordata, Nematoda and Arthropoda, and Fungi (Fig. 4).

The CBM20 part of the tree exhibits several characteristics already well-known from previous bioinformatics analyses [24,25]. These are especially the clustering of the SBDs from bacilli (found in CGTases), actinomycetes (in α-amylases), and fungi (in both α-amylases and glucoamylases). It seems that this reflection of taxonomy is indeed a feature of the evolution of the CBM20 module [24] because cyanobacteria also form a separate cluster, between laforins and the GH13 amylopullulanases (Fig. 4). This trend is supported by four CBM20 modules in GH77 4-α-glucanotransferases, of which the three plant members clustered separately from the bacterial one. Remarkably CBM20 of laforin grouped with SBD from the *Thermomonospora curvata* α-amylase. This is most interesting because *T. curvata* CBM20 exhibits all sequence features of a real SBD [66] although it appears away from the other CBM20 modules of actinomycetes [25]. With regard to the large cluster of SBDs from *Bacillus* CGTases, the positions of the modules from *Bacillus agaradhaerens* (cgtBacag, upBacag) indicate a slightly different phylogeny (Fig. 4) in accordance with previous findings based on entire CGTase sequences [67]. The sole representative of family GH31, CBM20 of 6-α-glucosyltransferase from actinobacterium *Arthrobacter globiformis* [68] grouped with the SBDs present in proteobacteria, two in *Pseudomonas* and one in *Klebsiella*. The former enzymes are maltotetraose-forming exo-amylases of GH13 and the latter is described as an intermediate between these four-domain hydrolases and five-domain transferases in GH13 [25]. Finally, there is one more novel CBM20 member observed in the α-glucan water dikinase from *Arabidopsis thaliana* [69], which interestingly is placed on a common branch with the module from the GH77 *Bacteoroides fragilis* 4-α-glucanotransferase, whereas the three plant 4-α-glucanotransferases are positioned separately adjacent to the borderline (Fig. 4).

The proposed joining of the two CBM20 and CBM21 families into one CBM clan raises a question about the possibility of the existence of an intermediate sequence. The modules from GH13 bacterial amylopullulanases [70–74] clustered most closely to the borderline and rather distant from the other clusters in the CBM20 part of the tree (Fig. 4). This module from amylopullulanase is therefore a candidate for an evolutionary intermediate between the two CBM families. This is in line with the presence of the module in the interior region of the domain organization as seen often in CBM21 (Fig. 1) and opposed to most CBM20 modules being either the N-terminal or the C-terminal domain.

As indicated in Experimental procedures, the most current update of the CAZy server contained 22 and six new members in CBM20 and CBM21, respectively, not present in Table 1. Of the 22 in CBM20, the added members were as follows: seven GH13 (four CGTases, two amylopullulanases, and one maltogenic α-amylase), six GH15 glucoamylases (four of them were from patents), one GH77 4-α-glucanotransferase, one genethonin-1 (from rat), five unknown proteins of animal origin (four from insect and one from fish), two carbohydrate esterases of the family CE-1 (both from Archaea), and one endoribonuclease E (from rice). With regard to the six recently added members in CBM21, five were putative protein phosphatases (or their regulatory subunits) and one was the unknown patented sequence from yeast, but there were no new amylolytic enzymes.

It is worth mentioning that the PSI-BLAST [75] searches using the above-mentioned added CBM sequences as queries revealed many new potential members of both CBM families. It is therefore reasonable to expect that in the future the number of members in the families in CAZy will continue to increase, as well as the spectrum of proteins with novel specificities. At present, in addition to the results shown in Fig. 4, the archaeal carbohydrate esterases of the CAZy CE-1 family [3], from *Pyrococcus furiosus* [76] and *Thermococcus kodakaraensis* [77], can be of special interest. Their CBM20 modules are most similar to those of GH13 amylopullulanases (possible intermediates between CBM20 and CBM21) included in the present study (Fig. 4). Moreover, and surprisingly, our PSI-BLAST searches clearly indicated that a similar CBM20 module is present in the GH13 (i.e., α-amylase family) branching enzymes (e.g. from *Equus caballus* [78]), which should also be included in the CAZy CBM20 classification.

Proposal for a new clan of CBM

Based on the bioinformatics analysis of SBD modules from CBM20 and CBM21 families, the hypothesis is proposed that the two types of real (functional) starch-binding domains, i.e., the C- and N-terminal SBDs thus far found in CBM20 and CBM21, respectively, share a common evolutionary origin. Because of this and the likelihood that CBM20 and CBM21 modules have similar secondary and tertiary structures, it is proposed to group the two SBD families, CBM20 and CBM21, into a hierarchically higher level of CAZy classification, i.e., a common CBM clan. An enzyme clan consists of a group of enzyme families with a common ancestry, very similar tertiary structure and conserved catalytic machinery and reaction mechanism [79]. Here we propose that a clan of carbohydrate-binding modules contains CBM families having a common evolutionary origin, similar tertiary structure and similar binding site residues, and mode of carbohydrate binding.

Experimental procedures

The set of analysed amino acid sequences of the CBM20 and CBM21 modules includes 181 proteins (Table 1). It was based on information in the CAZy server [3]. At the time of completing the sequence set (October 2004), there were 103 members of the CBM20 and 50 members of the CBM21 (Table 1). The last CAZy update (27 April 2005) contained an additional 22 and six members in CBM20 and CBM21, respectively. All of these sequences were subjected to PSI-BLAST searches [75].

Each SBD in the sequences studied was identified as follows: (a) for CBM20, the solved three-dimensional structures of the SBD from *Bacillus circulans* strain 251 CGTase [33] and *Aspergillus niger* glucoamylase [36,80] were used as templates; and (c) for CBM21, the best studied SBD from *Rhizopus oryzae* glucoamylase [49] was used as template. The exact position and length of the SBDs were, in all individual cases, supported by information extracted from the Pfam database [81] (Pfam Accession No. PF00686 for CBM20 and PF03370 for CBM21) as well as PSI-BLAST searches [75] using the default parameters.

All amino acid sequence alignments were performed using the program CLUSTALW [82] and then the alignments, where applicable, were manually adjusted. First, the sequences from CBM20 and CBM21 were aligned separately, starting with the sequences of amylolytic enzymes because of their mutual similarity. Second, the best conserved regions and residues [23,24], i.e., sequence fingerprints (625_TxxG, 640_LGxW, 661_PxW, and 689_WxxxxN; *B. circulans* strain 251 CGTase numbering including the 27-residue long signal peptide), were used in order to get the most reliable alignment of the CBM20 motifs. Finally, the same elements were applied for joining the two CBM families together into a final alignment, which was supported by the hydrophobic cluster analysis method [83].

The sequences were retrieved from GenBank [84] and UniProt [85]. The three-dimensional structures were taken from the PDB [86]. Secondary structures for the CBM21-type SBDs from *Lipomyces kononenkoae* α-amylase and *Rhizopus oryzae* glucoamylase were predicted using the GOR method [87,88] and SAM_T02 [89–91]. Fold recognition data for the CBM21-type SBD from *Rhizopus oryzae* glucoamylase and *Lipomyces kononenkoae* α-amylase were generated by the 3D-PSSM web server [61].

The evolutionary tree was calculated using the neighbour-joining method [92]. The Phylip format tree output was applied using the bootstrapping procedure [93]; the number of bootstrap trials used was 1000. The tree was drawn with the program TREEVIEW [94].

Acknowledgements

This work was financially supported by the VEGA grant no. 2/5067/5 from the Slovak Grant Agency for Science.

References

1 Horvathova V, Janecek S & Strudik E (2000) Amylolytic enzymes: their specificities, origins and properties. *Biologia (Bratisl)* **55**, 605–615.

2 Horvathova V, Janecek S & Strudik E (2001) Amylolytic enzymes: molecular aspects of their properties. *Gen Physiol Biophys* **20**, 7–32.

3 Coutinho PM & Henrissat B (1999) Carbohydrate-active enzymes: an integrated database approach. In *Recent Advances in Carbohydrate Bioengineering* (Gilbert HJ, Davies G, Henrissat B & Svensson B, eds), pp. 3–12. The Royal Society of Chemistry, Cambridge, UK.

4 Janecek S (2002) How many conserved sequence regions are there in the α-amylase family? *Biologia (Bratisl)* **57** (Suppl. 11), 29–41.

5 MacGregor EA, Janecek S & Svensson B (2001) Relationship of sequence and structure to specificity in the α-amylase family of enzymes. *Biochim Biophys Acta* **1546**, 1–20.

6 Zona R, Chang-Pi-Hin F, O'Donohue MJ & Janecek S (2004) Bioinformatics of the glycoside hydrolase family 57 and identification of catalytic residues in amylopullulanase from *Thermococcus hydrothermalis*. *Eur J Biochem* **271**, 2863–2872.

7 Frandsen TP & Svensson B (1998) Plant α-glucosidases of the glycoside hydrolase family 31. Molecular properties, substrate specificity, reaction mechanism, and comparison with family members of different origin. *Plant Mol Biol* **37**, 1–13.

8 Matsuura Y, Kusunoki M, Harada W & Kakudo M (1984) Structure and possible catalytic residues of Taka-amylase A. *J Biochem* **95**, 697–702.

9 Mikami B, Hehre EJ, Sato M, Katsube Y, Hirose M, Morita Y & Sacchettini JC (1993) The 2.0-Å resolution structure of soybean β-amylase complexed with α-cyclodextrin. *Biochemistry* **32**, 6836–6845.

10 Lovering AL, Lee SS, Kim YW, Withers SG & Strynadka NCJ (2005) Mechanistic and structural analysis of a family 31 α-glycosidase and its glycosyl-enzyme intermediate. *J Biol Chem* **280**, 2105–2115.

11 Aleshin A, Golubev A, Firsov LM & Honzatko RB (1992) Crystal structure of glucoamylase from *Aspergillus awamori* var. X100 to 2.2-Å resolution. *J Biol Chem* **267**, 19291–19298.

12 Imamura H, Fushinobu S, Yamamoto M, Kumasaka T, Jeon BS, Wakagi T & Matsuzawa H (2003) Crystal structures of 4-α-glucanotransferase from *Thermococcus litoralis* and its complex with an inhibitor. *J Biol Chem* **278**, 19378–19386.

13 Henrissat B & Davies G (1997) Structural and sequence-based classification of glycoside hydrolases. *Curr Opin Struct Biol* **7**, 637–644.

14 Rye CS & Withers SG (2000) Glycosidase mechanisms. *Curr Opin Chem Biol* **4**, 573–580.

15 Søgaard M, Kadziola A, Haser R & Svensson B (1993) Site-directed mutagenesis of histidine 93, aspartic acid 180, glutamic acid 205, histidine 290, and aspartic acid 291 at the active site and tryptophan 279 at the raw starch binding site in barley α-amylase 1. *J Biol Chem* **268**, 22480–22484.

16 Tibbot BK, Wong DWS & Robertson GH (2002) Studies on the C-terminal region of barley α-amylase 1 with emphasis on raw starch-binding. *Biologia (Bratisl)* **57** (Suppl. 11), 229–238.

17 Hostinova E, Solovicova A, Dvorsky R & Gasperik J (2003) Molecular cloning and 3D structure prediction of the first raw-starch-degrading glucoamylase without a separate starch-binding domain. *Arch Biochem Biophys* **411**, 189–195.

18 Kadziola A, Søgaard M, Svensson B & Haser R (1998) Molecular structure of a barley α-amylase-inhibitor complex: implications for starch binding and catalysis. *J Mol Biol* **278**, 205–217.

19 Robert X, Haser R, Gottschalk TE, Ratajczak F, Driguez H, Svensson B & Aghajari N (2003) The structure of barley α-amylase isozyme 1 reveals a novel role of domain C in substrate recognition and binding: a pair of sugar tongs. *Structure* **11**, 973–984.

20 Bozonnet S, Bønsager BC, Kramhøft B, Mori H, Abou Hachem M, Willemoës M, Jensen MT, Fukuda K, Nielsen PK, Juge N, Aghajari N, Tranier S, Robert X, Haser R & Svensson B (2005) Binding of carbohydrates and protein inhibitors to the surface of α-amylases. *Biologia (Bratisl)* **60** (Suppl. 16), 27–36.

21 Coutinho PM & Henrissat B (1999) The modular structure of cellulases and other carbohydrate-active enzymes: an integrated database approach. In *Genetics, Biochemistry and Ecology of Cellulose Degradation* (Ohmiya K, Hayashi K, Sakka K, Kobayashi Y, Karita S & Kimura T, eds), pp. 15–23. Uni Publishers Co., Tokyo, Japan.

22 Rodriguez-Sanoja R, Oviedo N & Sanchez S (2005) Microbial starch-binding domain. *Curr Opin Microbiol* **8**, 260–267.

23 Svensson B, Jespersen H, Sierks MR & MacGregor EA (1989) Sequence homology between putative raw-starch binding domains from different starch-degrading enzymes. *Biochem J* **264**, 309–311.

24 Janecek S & Sevcik J (1999) The evolution of starch-binding domain. *FEBS Lett* **456**, 119–125.

25 Janecek S, Svensson B & MacGregor EA (2003) Relation between domain evolution, specificity, and taxonomy of the α-amylase family members containing a C-terminal starch-binding domain. *Eur J Biochem* **270**, 635–645.

26 Juge N, Le Gal-Coeffet MF, Furniss CSM, Gunning AP, Kramhoeft B, Morris VJ, Williamson G & Svensson B (2002) The starch binding domain of glucoamylase from *Aspergillus niger*: overview of its structure, function, and role in raw-starch hydrolysis. *Biologia (Bratisl)* **57** (Suppl. 11), 239–245.

27 Boraston AB, Bolam DN, Gilbert HJ & Davies GJ (2004) Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J* **382**, 769–781.

28 Bibel M, Brettl C, Gosslar U, Kiegshaeuser G & Liebl W (1998) Isolation and analysis of genes for amylolytic enzymes of the hyperthermophilic bacterium *Thermotoga maritima*. *FEMS Microbiol Lett* **158**, 9–15.

29 Sumitani J, Tottori T, Kawaguchi T & Arai M (2000) New type of starch-binding domain: the direct repeat motif in the C-terminal region of *Bacillus* sp, 195 α-amylase contributes to starch binding and raw starch degrading. *Biochem J* **350**, 477–484.

30 Abe A, Tonozuka T, Sakano Y & Kamitori S (2004) 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* **335**, 811–822.

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

32 Klein C & Schulz GE (1991) Structure of cyclodextrin glycosyltransferase refined at 2.0 Å resolution. *J Mol Biol* **217**, 737–750.

33 Lawson CL, van Montfort R, Strokopytov B, Rozeboom HJ, Kalk KH, de Vries GE, Penninga D, Dijkhuizen L & Dijkstra BW (1994) Nucleotide sequence and X-ray structure of cyclodextrin glycosyltransferase from *Bacillus circulans* strain 251 in a maltose-dependent crystal form. *J Mol Biol* **236**, 590–600.

34 Knegtel RM, Wind RD, Rozeboom HJ, Kalk KH, Buitelaar RM, Dijkhuizen L & Dijkstra BW (1996) Crystal structure at 2.3 Å resolution and revised nucleotide sequence of the thermostable cyclodextrin glycosyltransferase from *Thermonanaerobacterium thermosulfurigenes* EM1. *J Mol Biol* **256**, 611–622.

35 Harata K, Haga K, Nakamura A, Aoyagi M & Yamane K (1996) X-Ray structure of cyclodextrin glucanotransferase from alkalophilic *Bacillus* sp. 1011. Comparison of two independent molecules at 1.8 Å resolution. *Acta Crystallogr* **D52**, 1136–1145.

36 Sorimachi K, Jacks AJ, Le Gal-Coeffet MF, Williamson G, Archer DB & Williamson MP (1996) Solution structure of the granular starch binding domain of glucoamylase from *Aspergillus niger* by nuclear magnetic resonance spectroscopy. *J Mol Biol* **259**, 970–987.

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

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

39 Goto M, Semimaru T, Furukawa K & Hayashida S (1994) Analysis of the raw starch-binding domain by mutation of a glucoamylase from *Aspergillus awamori* var. *kawachi* expressed in *Saccharomyces cerevisiae*. *Appl Environ Microbiol* **60**, 3926–3930.

40 Chen L, Coutinho PM, Nikolov Z & Ford C (1995) Deletion analysis of the starch-binding domain of *Aspergillus* glucoamylase. *Protein Eng* **8**, 1049–1055.

41 Penninga D, van der Veen BA, Knegtel RMA, van Hijum SAFT, Rozeboom HJ, Kalk KH, Dijkstra BW & Dijkhuizen L (1996) The raw starch binding domain of cyclodextrin glycosyltransferase from *Bacillus circulans* strain 251. *J Biol Chem* **271**, 32777–32784.

42 Sorimachi K, Le Gal-Coeffet MF, Williamson G, Archer DB & Williamson MP (1997) Solution structure of the granular starch binding domain of *Aspergillus niger* glucoamylase bound to β-cyclodextrin. *Structure* **5**, 647–661.

43 Giardina T, Gunning AP, Juge N, Faulds CB, Furniss CS, Svensson B, Morris VJ & Williamson G (2001) Both binding sites of the starch-binding domain of *Aspergillus niger* glucoamylase are essential for inducing a conformational change in amylose. *J Mol Biol* **313**, 1149–1159.

44 Dalmia BK, Schütte K & Nikolov ZL (1995) Domain E of *Bacillus macerans* cyclodextrin glucanotransferase: an independent starch-binding domain. *Biotechnol Bioeng* **47**, 575–584.

45 Ohdan K, Kuriki T, Takata H, Kaneko H & Okada S (2000) Introduction of raw starch-binding domains into *Bacillus subtilis* α-amylase by fusion with the starch-binding domain of *Bacillus* cyclomaltodextrin glucanotransferase. *Appl Environ Microbiol* **66**, 3058–3064.

46 Cornett CAG, Fang TY, Reilly PJ & Ford C (2003) Starch-binding domain shuffling in *Aspergillus niger* glucoamylase. *Protein Eng* **16**, 521–529.

47 Ji Q, Vincken JP, Suurs LCJM & Visser RGF (2003) Microbial starch-binding domains as a tool for targeting proteins to granules during starch biosynthesis. *Plant Mol Biol* **51**, 789–801.

48 Hua YW, Chi MC, Lo HF, Hsu WH & Lin LL (2004) Fusion of *Bacillus stearothermophilus* leucine aminopeptidase II with the raw-starch-binding domain of *Bacillus* sp. strain TS-23 α-amylase generates a chimeric enzyme with enhanced thermostability and catalytic activity. *J Ind Microbiol Biotechnol* **31**, 273–277.

49 Ashikari T, Nakamura N, Tanaka Y, Kiuchi N, Shibano Y, Tanaka T, Amachi T & Yoshizumi H (1986) *Rhizopus* raw-starch-degrading glucoamylase: its cloning and expression in yeast. *Agric Biol Chem* **50**, 957–964.

50 Bork P, Dandekar T, Eisenhaber F & Huynen M (1998) Characterization of targeting domains by sequence analysis: glycogen-binding domains in protein phosphatases. *J Mol Med* **76**, 77–79.

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

52 Janecek S (2002) A motif of a microbial starch-binding domain found in human genethonin. *Bioinformatics* **18**, 1534–1537.

53 Wang J, Stuckey JA, Wishart MJ & Dixon JE (2002) A unique carbohydrate binding domain targets the lafora disease phosphatase to glycogen. *J Biol Chem* **277**, 2377–2380.

54 Lohi HT & Minassian BA (2005) Starch-like polyglucosan formation in neuronal dendrites in the Lafora form of human epilepsy: a theory of pathogenesis. *Biologia (Bratisl)* **60** (Suppl. 16), 123–129.

55 Ceulemans H, Stalmans W & Bollen M (2002) Regulator-driven functional diversification of protein phosphatase-1 in eukaryotic evolution. *Bioessays* **24**, 371–381.

56 Armstrong CG, Doherty MJ & Cohen PT (1998) Identification of the separate domains in the hepatic glycogen-targeting subunit of protein phosphatase 1 that interact with phosphorylase *a*, glycogen and protein phosphatase 1. *Biochem J* **336**, 699–704.

57 Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A, Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M & Tabata S (2001) Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Res* **8**, 205–213.

58 Nakamura Y, Kaneko T, Sato S, Mimuro M, Miyashita H, Tsuchiya T, Sasamoto S, Watanabe A, Kawashima K, Kishida Y, Kiyokawa C, Kohara M, Matsumoto M, Matsuno A, Nakazaki N, Shimpo S, Takeuchi C, Yamada M & Tabata S (2003) Complete genome structure of *Gloeobacter violaceus* PCC 7421, a cyanobacterium that lacks thylakoids. *DNA Res* **10**, 137–145.

59 Wouters J, Bergman B & Janson S (2003) Cloning and expression of a putative cyclodextrin glucosyltransferase from the symbiotically competent cyanobacterium *Nostoc* sp. PCC 9229. *FEMS Microbiol Lett* **219**, 181–185.

60 Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M & Tabata S (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res* **3**, 109–136.

61 Kelley LA, MacCallum RM & Sternberg MJE (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol* **299**, 499–520.

62 Dauter Z, Dauter M, Brzozowski AM, Christensen S, Borchert TV, Beier L, Wilson KS & Davies GJ (1999) X-ray structure of Novamyl, the five-domain 'maltogenic' α-amylase from *Bacillus stearothermophilus*: maltose and acarbose complexes at 1.7 Å resolution. *Biochemistry* **38**, 8385–8392.

63 Steyn AJ, Marmur J & Pretorius IS (1995) Cloning, sequence analysis and expression in yeasts of a cDNA containing a *Lipomyces kononenkoae* α-amylase-encoding gene. *Gene* **166**, 65–71.

64 Coutinho PM & Reilly PJ (1997) Glucoamylase structural, functional, and evolutionary relationships. *Proteins* **29**, 334–347.

65 Williamson MP, Le Gal-Coeffet MF, Sorimachi K, Furniss CS, Archer DB & Williamson G (1997) Function of conserved tryptophans in the *Aspergillus niger* glucoamylase 1 starch binding domain. *Biochemistry* **36**, 7535–7539.

66 Janda L, Damborsky J, Petricek M, Spizek J & Tichy P (2000) Molecular characterization of the *Thermomonospora curvata* aglA gene encoding a thermotolerant α-1,4-glucosidase. *J Appl Microbiol* **88**, 773–783.

67 Martins RF, Delgado O & Hatti-Kaul R (2003) Sequence analysis of cyclodextrin glycosyltransferase from the alkaliphilic *Bacillus agaradhaerens*. *Biotechnol Lett* **25**, 1555–1562.

68 Mukai K, Maruta K, Satouchi K, Kubota M, Fukuda S, Kurimoto M & Tsujisaka Y (2004) Cyclic tetrasaccharide-synthesizing enzymes from *Arthrobacter globiformis* A19. *Biosci Biotechnol Biochem* **68**, 2529–2540.

69 Baunsgaard L, Lutken H, Mikkelsen R, Glaring MA, Pham TT & Blennow A (2005) A novel isoform of glucan, water dikinase phosphorylates pre-phosphorylated α-glucans and is involved in starch degradation in *Arabidopsis*. *Plant J* **41**, 595–605.

70 Mathupala S, Saha BC & Zeikus JG (1990) Substrate competition and specificity at the active site of amylopullulanase from *Clostridium thermohydrosulfuricum*. *Biochem Biophys Res Commun* **166**, 126–132.

71 Melasniemi H, Paloheimo M & Hemio L (1990) Nucleotide sequence of the α-amylase-pullulanase gene from *Clostridium thermohydrosulfuricum*. *J Gen Microbiol* **136**, 447–454.

72 Lee SP, Morikawa M, Takagi M & Imanaka T (1994) Cloning of the *aapT* gene and characterization of its product, α-amylase-pullulanase (AapT), from thermophilic and alkaliphilic *Bacillus* sp. strain XAL601. *Appl Environ Microbiol* **60**, 3764–3773.

73 Sahm K, Matuschek M, Mueller H, Mitchell WJ & Bahl H (1996) Molecular analysis of the *amy* gene locus of *Thermoanaerobacterium thermosulfurigenes* EM1 encoding starch-degrading enzymes and a binding protein-dependent maltose transport system. *J Bacteriol* **178**, 1039–1046.

74 Chen JT, Chen MC, Chen LL & Chu WS (2001) Structure and expression of an amylopullulanase gene from *Bacillus stearothermophilus* TS-23. *Biotechnol Appl Biochem* **33**, 189–199.

75 Altschul SF, Stephen F, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.

76 Robb FT, Maeder DL, Brown JR, DiRuggiero J, Stump MD, Yeh RK, Weiss RB & Dunn DM (2001) Genomic sequence of hyperthermophile, *Pyrococcus furiosus*: implications for physiology and enzymology. *Methods Enzymol* **330**, 134–157.

77 Fukui T, Atomi H, Kanai T, Matsumi R, Fujiwara S & Imanaka T (2005) Complete genome sequence of the hyperthermophilic archaeon *Thermococcus kodakaraensis* KOD1 and comparison with *Pyrococcus* genomes. *Genome Res* **15**, 352–363.

78 Ward TL, Valberg SJ, Adelson DL, Abbey CA, Binns MM & Mickelson JR (2004) Glycogen branching enzyme (GBE1) mutation causing equine glycogen storage disease IV. *Mamm Genome* **15**, 570–577.

79 Henrissat B & Bairoch A (1996) Updating the sequence-based classification of glycosyl hydrolases. *Biochem J* **316**, 695–696.

80 Tanaka Y, Ashikari T, Nakamura N, Kiuchi N, Shibano Y, Amachi T & Yoshizumi H (1986) Comparison of amino acid sequences of three glucoamylases and their structure-function relationships. *Agric Biol Chem* **50**, 965–969.

81 Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M & Sonnhammer EL (2002) The Pfam protein families database. *Nucleic Acids Res* **30**, 276–280.

82 Thompson JD, Higgins DG & Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–4680.

83 Callebaut I, Labesse G, Durand P, Poupon A, Canard L, Chomilier J, Henrissat B & Mornon JP (1997) Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell Mol Life Sci* **53**, 621–645.

84 Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J & Wheeler DL (2004) GenBank: update. *Nucleic Acids Res* **32**, D23–D26.

85 Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N & Yeh LS (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32**, D115–D119.

86 Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD & Zardecki C (2002) The protein data bank. *Acta Crystallogr* **D58**, 899–907.

87 Garnier J, Gibrat JF & Robson B (1996) GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol* **266**, 540–553.

88 Combet C, Blanchet C, Geourjon C & Deleage G (2000) NPS@: Network Protein Sequence Analysis. *Trends Biochem Sci* **25**, 147–150.

89 Karplus K, Karchin R, Barrett C, Tu S, Cline M, Diekhans M, Grate L, Casper J & Hughey R (2001) What is the value added by human intervention in protein structure prediction? *Proteins* **45** (S5), 86–91.

90 Karchin R, Cline M, Mandel-Gutfreund Y & Karplus K (2003) Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins* **51**, 504–514.

91 Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M & Hughey R (2003) Combining local-structure, fold-recognition, and new-fold methods for protein structure prediction. *Proteins* **53** (S6), 491–496.

92 Saitou N & Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406–425.

93 Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791.

94 Page RD (1996) TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci* **12**, 357–358.