Sphingomonas paucimobilis β -glucosidase Bgl1: a member of a new bacterial subfamily in glycoside hydrolase family 1

Ana Rita MARQUES, Pedro M. COUTINHO, Paula VIDEIRA, Arsénio M. FIALHO and Isabel SÁ-CORREIA¹

Centro de Engenharia Biológica e Química, Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisboa, Portugal

The *Sphingomonas paucimobilis* β -glucosidase Bgl1 is encoded by the *bgl1* gene, associated with an 1308 bp open reading frame. The deduced protein has a potential signal peptide of 24 amino acids in the N-terminal region, and experimental evidence is consistent with the processing and export of the Bgl1 protein through the inner membrane to the periplasmic space. A His₆-tagged 44.3 kDa protein was over-produced in the cytosol of *Escherichia coli* from a recombinant plasmid, which contained the *S. paucimobilis* *bgl1* gene lacking the region encoding the putative signal peptide. Mature β -glucosidase Bgl1 is specific for aryl- β -glucosides and has no apparent activity with oligosaccharides derived from cellulose hydrolysis and other saccharides. A structure-based alignment established structural relations between *S. paucimobilis* Bgl1 and other members of the

glycoside hydrolase (GH) family 1 enzymes. At subsite –1, the conserved residues required for catalysis by GH1 enzymes are present in Bgl1 with only minor differences. Major differences are found at subsite +1, the aglycone binding site. This alignment seeded a sequence-based phylogenetic analysis of GH1 enzymes, revealing an absence of horizontal transfer between phyla. Bootstrap analysis supported the definition of subfamilies and revealed that Bgl1, the first characterized β -glucosidase from the genus *Sphingomonas*, represents a very divergent bacterial subfamily, closer to archaeal subfamilies than to others of bacterial origin.

Key words: *bgl1* gene, CAZy database, β -glucosidase, hydrolase family 1, periplasmic proteins, *Sphingomonas paucimobilis*.

INTRODUCTION

β -Glucosidases (β -glucoside glucohydrolases; E.C. 3.2.1.21) cleave β -glycosidic bonds between two or more carbohydrates, or between a carbohydrate and a non-carbohydrate moiety, by a double displacement mechanism [1,2]. They may act on a broad range of β -D-glycosides and play differentiated and important biological roles in eukaryotic and prokaryotic organisms, ranging from oligo- or polysaccharide hydrolysis in micro-organisms [3,4], to plant resistance to phytopathogens [5]. β -Glucosidase activities are found in two distinct families of glycoside hydrolases [6,7], described in the CAZy (Carbohydrate-Active enZymes) database (<http://afmb.cnrs-mrs.fr/CAZY>), which is dedicated to family classification of glycoside hydrolases and other carbohydrate-active enzymes [8]. Glycoside hydrolase (GH) family 1 exhibits the (α/β)₈-barrel-fold characteristic of clan GH-A, the largest clan covering mostly β -glycan-retaining glycosidases [9]. Different structures of archaeal, bacterial and plant origin provide insight into both structure and catalysis in the GH1 family. In clan GH-A, two glutamates act as catalytic acids and catalytic nucleophiles in a retention mechanism involving two steps: (i) the formation of the glucosyl–enzyme complex with concomitant aglycone departure (glycosylation step) and (ii) hydrolysis of the glucosyl–enzyme complex by a water molecule (the deglycosylation step) [1,2]. A few GH1 enzyme complexes, corresponding to the characteristic enzyme-bound reaction intermediate {Protein Data Bank (PDB) structure 1E4I and [11]}, define the glycone subsite, or subsite –1 in the usual nomenclature [12]. Recent enzyme–substrate complexes [13,14] provide a first insight of the aglycone site, or subsite +1. The systematic identification of sequences, coupled with the rapid increase in genomic analysis, has allowed the classification of more than 250 proteins in the

GH1 family, although a significant number have yet to be characterized biochemically.

In the present study, we report the identification and sequence analysis of the *bgl1* gene from *Sphingomonas paucimobilis* ATCC 31461 and the biochemical characterization of the encoded protein, a periplasmic β -glucosidase. This bacterial strain was isolated from plant tissue [15]. A number of *Sphingomonas* species associate with plants, some being phytopathogenic [16–18]. *S. paucimobilis* ATCC 31461 synthesizes high levels of the commercial exopolysaccharide gellan gum, which is used as a gelling agent in foods and microbiological culture media. Previous studies from this laboratory have identified and characterized the genes and enzymes involved in gellan biosynthesis [19–21].

In the present study, the *bgl1* gene, identified in our *S. paucimobilis* ATCC 31461 gene bank [19], is shown to encode the divergent periplasmic aryl- β -glucosidase Bgl1, the first β -glucosidase from the genus *Sphingomonas* to be characterized. Complementary biochemical and *in silico* analysis defined its cellular location, and clearly positioned Bgl1 within the > 250 complete sequences belonging to family GH1. Multisequence alignment seeded a phylogenetic analysis that correlated significantly with the taxonomic analysis, and clearly revealed that Bgl1 is representative of a new GH1 bacterial subfamily, closer to archaeal subfamilies.

MATERIALS AND METHODS

Strains, plasmids and growth conditions

The strains and plasmids used in this work are listed in Table 1. *Escherichia coli* strain XL1-Blue was used as the host for most of

Abbreviations used: CAZy, carbohydrate-active enZymes; DOE-JGI, Department of Energy Joint Genome Institute; GH, glycoside hydrolase; IPTG, isopropyl β -D-thiogalactoside; ORF, open reading frame; PDB, Protein Data Bank; pNPG, *p*-nitrophenyl β -D-glucopyranoside.

¹ To whom correspondence should be addressed (e-mail isacorreia@ist.utl.pt).

The nucleotide sequence data reported will appear in DDBJ, EMBL, GenBank® and GSDB Nucleotide Sequence Databases under the accession number AF305688 (or Spauc_bgl_AAG59862).

Table 1 Bacterial strains and plasmids used in this workAp^r, ampicillin resistance; Cm^r, chloramphenicol resistance; Km^r, kanamycin resistance; Tc^r, tetracycline resistance.

Strains and plasmids	Genotype or phenotype	Source and/or reference
Strains		
<i>Escherichia coli</i>		
XL1-Blue	<i>recA1 lac</i> [F' <i>proAB lacI</i> ^q Δ M15 Tn10 (Tc ^r), <i>thi</i>	Stratagene®
BL21(DE3)	F' <i>ompT hsdS₈(r₆⁻m₈⁻) gal dcm</i> (DE3)	Novagen®
SURE	[e14 ⁻ (mcrA) Δ (mcrCB-hsdSMR-mrr)171 endA1 supE44 thi-1 gyrA96 relA1 <i>lac recB recJ sbsC umuC::Tn5</i> (Km ^r) <i>uvrC</i> [F' <i>proAB lacI</i> ^q Δ M15 Tn10 (Tc ^r)]	Stratagene®
Plasmids		
Hypercos 1	8.8 kb, Ap ^r , double <i>cos</i> , T3, T7	A. Pühler*
pBC SK	3.4 kb, Cm ^r , <i>lacZ</i> , T3, T7	Stratagene®
pET29(b)	5.4 kb, Km ^r , <i>lacI</i> , S-tag, His-tag (optional)	Novagen®
PWH844	pQE-9 derivative, <i>lacI</i> ^q , 5.0 kb, Ap ^r , T5, His-tag	[24]
pC4	Cosmid from <i>S. paucimobilis</i> ATCC 31461 genomic library in Hypercos 1; 35 kb insert	Present work
pGlul	pBC SK derivative with the 1.6 kb <i>Xma</i> III fragment from pC4, containing <i>bglI</i> gene inserted	Present work
PBgl1 _c	pET29(b) derivative containing a <i>Eco</i> RI- <i>Hind</i> III fragment amplified from pGlul using primers A and B (Figure 1)	Present work
pBgl1 _p	pWH844 derivative containing a <i>Pst</i> I- <i>Hind</i> III fragment amplified from pGlul using primers C and D (Figure 1)	Present work

* University of Bielefeld, Faculty of Biology, Bielefeld, Germany.

the cloning procedures and it was grown in Lennox broth (Sigma Chemical Co, St. Louis, MO, U.S.A.) at 37 °C. *E. coli* BL21 (DE3) and *E. coli* SURE were used as hosts for the over-expression of total or partial *bglI* sequences from pET29(b)- and pWH844-derived plasmids respectively. These cloning translational vectors were used to over-express the complete (c) and the partial (p; lacking the region coding for a putative signal peptide) *bglI* sequence respectively. For protein over-expression, cells of *E. coli* BL21 (DE3) harbouring pBgl1_c or *E. coli* SURE harbouring pBgl1_p, were grown in Lennox broth supplemented with 0.5 % (w/v) glucose or in Super Broth (32 g/l tryptone, 20 g/l yeast extract, 5 g/l NaCl) respectively. When required, antibiotics were added to the selective media at the following concentrations: 150 mg/l ampicillin, 100 mg/l kanamycin and 25 mg/l chloramphenicol.

DNA manipulations and reagents

Plasmid DNA was purified from *E. coli* cultures by the alkaline lysis procedure [22] or with a QIAprep spin plasmid kit (Qiagen, Santa Clarita, CA, U.S.A.). DNA restriction, agarose gel electrophoresis, cloning procedures and transformation of *E. coli* were carried out following established protocols [23]. The oligonucleotides were purchased from Amersham Biosciences (Carnaxide, Portugal). Restriction enzymes were from Gibco-BRL (Life Technologies, Barcelona, Spain), and *Pwo* polymerase was from Roche (Mannheim, Germany). All the other chemicals were from Sigma Chemical Co.

Cloning of total and partial *bglI* sequences

Computer analysis of the partial nucleotide sequence of cosmid pC4, belonging to the *S. paucimobilis* ATCC 31461 gene bank described previously [19,20], revealed the presence of an open reading frame (ORF), named the *bglI* gene. A *Xma*III fragment of pC4 (1619 bp), containing the entire *bglI* gene, was cloned in pBC SK, generating pGlul, which was used for *bglI* sequencing. For the heterologous over-expression of the *bglI* gene, the

complete sequence was amplified by PCR using *Pwo* polymerase, oligonucleotides A and B (Figure 1) and plasmid pGlul as the template. The resulting PCR product (1324 bp), after endonuclease hydrolysis, was cloned into *Eco*RI-*Hind*III sites of pET29(b) under the control of T7 RNA polymerase promoter, generating pBgl1_c. This recombinant plasmid carried an S-tag in the 5' end, followed by the complete *bglI* gene (without the stop codon) and the coding sequence for six histidine residues at the 3' end (Figure 1). To express the *bglI* sequence lacking the region coding for the predicted signal peptide (Figure 1), which is under the control of the T5 promoter, plasmid pBgl1_p was prepared by inserting the 1251 bp PCR product, obtained by using the oligonucleotides C and D (Figure 1), into the *Pst*I and *Hind*III sites of the translational vector pWH844 [24]. The resulting plasmid, pBgl1_p, carried the *bglI* gene without the first 75 nucleotides preceded by the sequence coding for the six histidine residues from the pWH844 vector.

Over-expression and purification of complete and partial *bglI* in *E. coli*

To over-express the complete *S. paucimobilis bglI* gene, transformants of *E. coli* BL21(DE3) harbouring plasmid pBgl1_c [or the cloning vector pET29(b) as the control] were cultivated in 100 ml of Lennox broth, containing 0.5 % (w/v) glucose and 50 mg/l of kanamycin, at 37 °C until the attenuation (D_{640}) of the culture reached 0.6 ± 0.1 . At this point, *bglI* transcription was induced by the addition of 0.1 mM isopropyl β -D-thiogalactoside (IPTG), followed by an additional period of 3 h of cultivation. To obtain the periplasmic fraction, induced cells were resuspended in osmotic shock buffer [20 % (w/v) sucrose, 200 mM Tris/HCl, 10 mM EDTA, 500 g/ml lysozyme] and incubated at room temperature for 15 min. After the addition of an equal volume of cold water, the cells were centrifuged at 13000 g for 15 min and the supernatant was recovered as the periplasmic fraction [25]. The buffer was replaced with 1 mM

Figure 1 Schematic representation of genetic manipulations carried out to prepare the recombinant plasmids used to express the complete (c) or the partial (p) *bgl1* gene sequence

(A) Hydrophobicity plot of the predicted Bgl1 sequence [28]. (B) Location of the putative signal peptide within Bgl1. (C) *bgl1* ORF showing the positions of the primers (arrows) used to generate the two Bgl1 constructions (double-headed arrows) used in this study. (D) Nucleotide sequences of the primers used in this work. The underlined sections indicate the target sequences for *EcoRI* (A), *HindIII* (B, D) and *PstI* (C) and the position of the primers is given using the thymidine of the initiation codon (TTG) of *bgl1* as base 1.

sodium phosphate buffer (pH 7.4) containing 50 mM NaCl and 10 mM imidazole, using Biomax 10 kDa membranes (Millipore Corporation, Bedford, MA, U.S.A.). To obtain the total crude cell extract, cells were resuspended in cold 1 mM sodium phosphate buffer (pH 7.4) containing 50 mM NaCl and 10 mM imidazole, disrupted by sonication (VibraCell; Sonics Material, Danbury, CT, U.S.A.) and centrifuged at 20000 *g* at 4 °C for 1 h. To over-express the *S. paucimobilis bgl1* gene without the region coding for the putative signal peptide, transformants of *E. coli* SURE carrying the plasmid pBgl1_p (or the translational vector pWH844 as the control) were cultivated at 37 °C in 100 ml of Super Broth with 150 mg/l ampicillin, until the attenuation (D_{640}) of the culture reached 0.6 ± 0.1 . The cells were then induced with IPTG. Crude cell extract was obtained by sonication, as described above. The overproduced His₆-tagged Bgl1 was purified in one step, under native conditions, following the manufacturer's recommendations (Amersham Biosciences). The fractions were applied to the HisTrap[®] column (Amersham Biosciences), which was pre-equilibrated with equilibration buffer [1 mM sodium phosphate buffer (pH 7.4), 50 mM NaCl and 10 mM imidazole]. In order to remove any unbound proteins, two washes with wash buffer [1 mM sodium phosphate buffer (pH 7.4), 50 mM NaCl and 100 mM imidazole] were performed.

The His₆-Bgl1 proteins were eluted with elution buffer [1 mM sodium phosphate buffer (pH 7.4), 50 mM NaCl and 200 mM imidazole]. The protein from the eluted fractions was transferred into 50 mM sodium phosphate buffer (pH 6.2) using Biomax 10 kDa membranes (Millipore). Proteins were concentrated in the Biomax membranes and used immediately for enzyme assays, or stored at 4 °C.

The proteins present in the cell crude extracts, the periplasmic fraction (30 μ g), or in the purified samples (2 μ g), were separated by SDS/PAGE using 10 % or 12.5 % (w/v) polyacrylamide gels. Low-molecular-mass standards (Amersham Biosciences) were used to estimate the molecular mass of the Bgl1 proteins overproduced with the two different genetic constructions. Gels were stained with Coomassie Brilliant Blue.

Enzyme assays, kinetic properties and substrate specificity of mature Bgl1

β -Glucosidase activity was determined using *p*-nitrophenyl β -D-glucopyranoside (*p*NPG) as the substrate. Assay mixtures (0.6 ml) containing 5 mM *p*NPG in 50 mM sodium phosphate buffer (pH 6.2) were incubated with the enzyme samples at 37 °C. The reactions were stopped by adding 0.5 ml of 1 M

Na_2CO_3 and the amount of *p*-nitrophenol produced was calculated assuming a molar absorption coefficient of $7400 \text{ l M}^{-1} \cdot \text{cm}^{-1}$ at 420 nm. Temperature and pH activity profiles were determined using *p*NPG (5 mM) as the substrate. Temperature-dependence was established in 50 mM sodium phosphate buffer (pH 6.2) and the thermal stability of the enzyme was determined by measuring the residual activity of the enzyme after 4 h of incubation at varying temperatures. pH-dependence was evaluated in 50 mM sodium acetate (pH 3.4–7.3) at 37 °C, and pH-stability was determined by measuring the residual activity of the enzyme after 24 h of incubation at 20 °C at varying pH levels. Enzyme activities were calculated based on the initial linear rates of product formation after subtraction of endogenous activity (measured in the enzyme assay mixture lacking the substrate). Control assays lacking the crude extracts/purified enzyme were also carried out. All the enzyme assays were performed in triplicate. Protein concentration was determined by the Bradford method [26], using BSA fraction V as the standard. One unit of enzyme activity corresponded to the release of $1 \mu\text{mol}$ of *p*-nitrophenol per min, under the assay conditions described above.

The kinetic constants, K_m and $V_{\max}$, corresponding to β -D-glucosidase activity of the purified mature BglI for the different *p*-nitrophenyl β -glycosides examined, arbutin, salicin and the various saccharides, were calculated using the Michaelis–Menten equation directly; the rectangular hyperbolic function was solved iteratively using Solver from Microsoft Excel. The k_{cat} values (min^{-1}) were calculated by dividing $V_{\max}$ (μmol of enzyme reaction product $\text{min}^{-1} \cdot \text{mg}^{-1}$) by 0.0226 to convert protein concentration into molarity, on the basis of a molecular mass of 43.1 kDa for mature BglI. β -Glucosidase activity for the different substrates was assayed by measuring either the nitrophenol or glucose that was released (UV Glucose Kit; Roche).

Sequence analysis

The purified recombinant plasmid pGluI, which carries a 1619 bp DNA fragment derived from pC4 cosmid and encompasses the *bglI* gene, was used as the template for DNA sequencing. Sequence data were analysed with DNASIS 3.0 software (Hitachi Software engineering, Brisbane, CA, U.S.A.). Database searches were performed with gapped BLAST [27] at the National Center for Biotechnology Information (NCBI) server (Bethesda, MD, U.S.A.) and against the curated sequences of CAZy (AFMB-CNRS, Marseille, France, at <http://afmb.cnrs-mrs.fr/CAZY>), where segments not corresponding to catalytic domains were excised from analysis. The presence of two BglI homologues was also detected by BLASTP analysis of all ORFs present in the *Novosphingobium aromaticivorans* (previously *Sphingomonas aromaticivorans*) genome (version 04dec00), whose data have been provided freely by the U.S. Department of Energy Joint Genome Institute (DOE-JGI; <http://www.jgi.doe.gov>) for use in this publication only. The presence and location of the coding sequences for the signal peptides in *BglI* and on the two homologues identified in the *N. aromaticivorans* genome were estimated using the Neural Networks and Hidden Markov models trained on Gram-negative bacteria using SignalP2.0 (BioCentrum-DTU, Technical University of Denmark, Lyngby, Denmark; <http://www.cbs.dtu.dk/services/SignalP-2.0>). The potential cellular location BglI was predicted with Psort Prediction (version 6.4) for Gram-negative bacteria (Institute of Medical Science, University of Tokyo, Japan; <http://psort.nibb.ac.jp>). The hydropathic profile of BglI was performed using Kyte–Doolittle values [28] at the Protein Hydrophilicity/Hydrophobicity Search and Comparison Server (Weizmann

Institute of Science, Rehovot, Israel; http://bioinformatics.weizmann.ac.il/hydroph/plotfft_hydroph.html).

Structures and sequences

The available GH1 glycosidases deposited in the PDB, namely 1CBG, 2MYR, 1PBG, 1GOW, 1BGG, 1QVB, 1QOX and 1E1E, were analysed using STRIDE [29] to obtain a secondary structure description. The remaining sequences used in this study were collected from CAZy [8] and from the DOE-JGI server containing the released *N. aromaticivorans* ORFs. These sequences include all known and complete GH1 sequences present in public databases and referenced in CAZy at the time of this study (January 2002). The identification of the residues involved in the interaction with different substrates or analogues at subsites -1 and $+1$ was performed on PDB structures 1E56, 1E4I and 1E73 with LPC software, using $4.0 \text{ \AA}$ ($1 \text{ \AA} = 0.1 \text{ nm}$) as the maximum contact distance [30].

Multisequence alignment

Starting with the structure-based alignment 'ghf1', obtained from the HOMSTRAD database [31], a more complete alignment was constructed and edited using DCSE 3.4 [32], a sequence editor and analysis tool that allows simultaneous assignment of primary and secondary structure to protein sequences. The sequence of BglI, all the eight structures already known for GH1 proteins and a set of 24 sequences of GH1 enzymes representative of all the known activities of this family, were chosen. Based on the principles of hydrophobic cluster alignment [33], a multi-sequence alignment of all 33 sequences was manually constructed. The alignment was performed by maintaining the hydrophobic residues in a regular arrangement in helices and β -strands by homology to the existing three-dimensional models, and placing most of the insertions or deletions in non-structured regions in the protein, often associated with positioning of turns. A global sequence alignment with 253 sequences was performed by ClustalW [34] in profile mode against the 33 sequences of the manual alignment produced earlier.

Phylogenetic analysis

Phylogenetic analysis of the 253 GH1 proteins compared in this study was performed. The manually edited 33-sequence alignment served as a profile to align the remaining sequences by ClustalW 1.82, a multiple sequence alignment program. The resulting alignment was analysed by TREECON software [35], using the Neighbour-Joining method [36] and supported by 500 bootstrap steps [37]. Only the tree branches that were resolved above 55 % on the bootstrap analysis were displayed. Subfamilies were defined as the branches with three or more sequences starting from the lowest branch point, at a resolution of above 85 % following bootstrap. As the ground step for bootstrap analysis, the sequence of *A. thaliana* (GenBank® accession no. AC020580/AA051335 or Athal_orf_AAG51335) was chosen as the most divergent in the analysis and for being the only one which failed to align correctly in the profile alignment. It was confirmed, using manual methods, that this sequence did belong to the GH1 family (results not shown).

Nucleotide sequence accession number

The complete sequence of *bglI* has been deposited in the GenBank® Nucleotide Sequence Database under the accession number AF305688 (or Spauc_bgl_AAG59862).

RESULTS

Sequence analysis of the *bgl1* gene

Computer analysis of the 1619 bp insert of plasmid pGlu1, derived from *S. paucimobilis* ATCC 31461 gene-bank cosmid pC4, revealed the presence of a unique ORF, designated the *bgl1* gene. This 1308-nucleotide-long ORF has a GC content of 68 % and a high frequency of G or C in the third position, characteristic of *Sphingomonas* genes [19,20,38]. Analysis of the pGlu1 fragment revealed a TTG codon, preceded (6 nucleotides upstream) by a putative ribosome-binding site (5'-GGAGA-3'). Since no other in-frame putative start codons (ATG or GTG) were found downstream of the TTG identified, the sequence analysis suggested that this TTG codon might provide the translational start for the *bgl1* gene in *S. paucimobilis* ATCC 31461. No typical -10 and -35 regions were identified upstream of the TTG start codon. Downstream of the TGA stop codon, no sequence resembling a *rho*-independent transcriptional terminator could be identified. The protein deduced from the *bgl1* complete sequence is composed of 435 amino acid residues, with a calculated molecular mass of 46753 Da. The presence of a potential signal peptide of 24 amino acids at the N-terminal region was identified, presumably directing the protein to the periplasmic space (probability of 0.921, based on Psort Prediction analysis). The most likely cleavage site (probability of 0.998) is located between amino acid residues 24 and 25 (V²² RA↓APP²⁷), as given by SignalP 2.0 and by the hydrophobicity plot constructed with Kyte-Doolittle values (Figure 1). The mature protein would then begin at Ala²⁵, having 411 residues, a calculated molecular mass of 44310 Da and a predicted pI of 7.75. The processing and export through the inner membrane of Bgl1, as suggested by computer analysis, are consistent with the results described below, showing that the mature protein can be purified from the *E. coli* periplasmic fraction following its heterologous over-expression.

A gapped BLAST search against the NCBI [27] non-redundant set of protein sequences revealed several GH1 β -glucosidases that were clearly homologous to Bgl1, the highest degree of

Table 2 Purification of *S. paucimobilis* ATCC 31461 mature β -glucosidase Bgl1-His₆ fusion protein

E. coli SURE cells were induced with IPTG to over-express pBgl1p (see Figure 2C, lane D). Enzyme activity was determined under the assay conditions described in the Materials and methods section, using pNPG as the substrate.

Purification step	Protein (g)	Total activity (units)	Specific activity (units/g)	Yield (%)	Purification (fold)
Bgl1 in crude extract	1.13	1132	1000	100.0	
His ₆ -tag purification	0.47	409.1	8760	36.1	8.8

homology (only 39 % identity and 52 % similarity) being found with β -glucosidase from *Streptomyces rochei* (TrEMBL:Q55000 or Sroch_bgl_Q55000). The analysis of the *Novosphingobium aromaticivorans* genome revealed two GH1 sequences with a putative signal peptide: gene 18 from contig 145 (Narom_orf_145.18) and gene 1 from contig 61 (Narom_orf_61.1), which exhibited 57 % and 55 % identity with Bgl1 respectively.

Heterologous overproduction and purification of the full and mature Bgl1 protein

When the complete *bgl1* gene sequence was over-expressed in IPTG-induced *E. coli* BL21(DE3) cells transformed with plasmid pBgl1_c (Figure 1), the single His₆-tagged protein purified by affinity chromatography from the periplasmic fraction was approx. 45 kDa (Figure 2A); this is consistent with the predicted size of the putative mature His₆-tagged protein lacking the N-terminal S-tag (encoded by the cloning vector) and the first 24 amino acids (44.3 kDa from the mature protein plus 1.3 kDa corresponding to the His₆-tag). However, the whole crude extract exhibited a predominant His₆-tagged protein of approx. 53 kDa in addition to the His₆-tagged 45 kDa protein, which was present in very low amounts (Figure 2B). A molecular mass of 53 kDa corresponds with the predicted size for the fusion protein S-Bgl1-His₆, including the signal peptide sequence (3.7 kDa for the N-terminal S-tag and the polylinker region, 46.7 kDa for the predicted complete Bgl1 amino acid sequence, plus 1.3 kDa for the histidine-tag). These results are consistent with the computational prediction of the processing and periplasmic location for the mature Bgl1 protein.

The β -glucosidase-specific activity (towards pNPG) detected in the purified crude extract prepared from BL21(DE3) cells transformed with pBgl1_c (Figure 2B, lane D) was very low (44 units/g of protein), suggesting that Bgl1 is not fully functional before maturation. As reported for other periplasmic enzymes [39], the purification yield of mature Bgl1 protein present in the periplasm of recombinant *E. coli* cells was very low. Therefore in order to characterize the mature Bgl1 protein, plasmid pBgl1_p was used to overproduce the *S. paucimobilis* mature Bgl1 protein, lacking the putative signal peptide, in the cytosol of recombinant *E. coli* cells. Construction with plasmid pWH844 avoided the presence of an S-tag in the recombinant protein whilst maintaining a His₆-tag for purification. The fused His₆-Bgl1 protein lacking the putative signal peptide was over-expressed in *E. coli* SURE from plasmid pBgl1_p (Figure 2C), and purified to homogeneity in a single step (Table 2), as judged by SDS/PAGE (Figure 2C, lane D). The specific activity of the purified cytosolic 45 kDa recombinant protein, putatively identical with the mature *S. paucimobilis* periplasmic Bgl1 protein, exhibited a high β -glucosidase-specific activity (1000 units/g of protein towards pNPG).

Figure 2 SDS/PAGE analysis of the purification of His₆-tagged Bgl1 proteins

Coomassie Blue-stained proteins separated by SDS/PAGE with 10 % (A, C) or 12.5 % (B) polyacrylamide during the purification of the His₆-tagged Bgl1 proteins produced in IPTG-induced *E. coli* BL21(DE3) cells harbouring plasmid pBgl1_c (A, B) or in *E. coli* SURE cells harbouring plasmid pBgl1_p (C). (A) Lane A, molecular markers; lane B, crude periplasmic fraction from *E. coli* BL21 (DE3) cells harbouring plasmid pBgl1_c (30 μ g of protein); lane C, Bgl1 protein resulting from His-tag purification from the periplasmic fraction in lane B (2 μ g of protein). (B) Lane A, molecular markers; lanes B and C, whole crude extracts (30 μ g of protein) prepared from *E. coli* BL21(DE3) cells harbouring the cloning vector pET29(b) and plasmid pBgl1_c respectively; lane D, Bgl1 proteins (2 μ g of protein) resulting from His-tag purification from the crude extract in lane C. (C) Lane A, molecular markers; lanes B and C, whole crude extract (30 μ g of protein) prepared from *E. coli* SURE cells harbouring the cloning vector pWH844 and plasmid pBgl1_p respectively; lane D, Bgl1 protein (2 μ g of protein) resulting from His₆-tag purification from the crude extract in lane C.

Table 3 Kinetic parameters of *S. paucimobilis* ATCC 31461 mature β -glucosidase Bgl1-His₆ fusion protein

E. coli SURE cells were induced with IPTG to over-express pBgl1p (see Figure 2C, lane D). k_{cat} values were calculated on the basis of His₆-Bgl1 as a 45-kDa monomer. ND, not detectable.

Substrate	K_m (mM)	V_{max} ($\mu\text{mol} \cdot \text{min}^{-1} \text{mg}^{-1}$)	k_{cat} (min^{-1})	k_{cat}/K_m ($\text{min}^{-1} \text{mM}^{-1}$)
Synthetic glycosides				
Aryl-glycosides				
<i>p</i> -Nitrophenyl β -D-glucopyranoside	1.3 ± 0.19	16.6 ± 3.07	734.5 ± 135.8	571.3 ± 134.0
<i>p</i> -Nitrophenyl β -D-galactopyranoside	6.5 ± 0.9	25.44 ± 6.54	1126 ± 289.6	175.7 ± 52.2
<i>p</i> -Nitrophenyl β -D-xylopyranoside	2.5 ± 0.35	2.6 ± 0.19	115.3 ± 8.6	46.2 ± 7.3
<i>p</i> -Nitrophenyl β -D-glucuronide	ND	—	—	—
<i>p</i> -Nitrophenyl β -D-mannopyranoside	ND	—	—	—
<i>p</i> -Nitrophenyl α -D-rhamnopyranoside	ND	—	—	—
<i>o</i> -Nitrophenyl β -D-glucopyranoside	8.3 ± 1.5	36.4 ± 0.63	1611 ± 28.1	259.2 ± 21.4
Natural glycosides				
Aryl glucosides				
Arbutin	7.6 ± 0.4	17.9 ± 1.32	791.15 ± 58.4	104.4 ± 9.5
Salicin	30.9 ± 1.4	9.2 ± 0.16	406.0 ± 6.9	13.2 ± 0.6
Saccharides				
Cellobiose	ND	—	—	—
Cellotetraose	ND	—	—	—
Lactose	ND	—	—	—

Figure 3 Saturation curves of purified His₆-Bgl1

Saturation curves of purified His₆-Bgl1 with (A) varying concentrations of arbutin (●) or salicin (○), and (B) nitrophenyl β -D-glucopyranosides substrates [pNPG (□); *o*-nitrophenyl β -D-glucopyranoside (■); *p*-nitrophenyl β -D-galactopyranoside (△); *p*-nitrophenyl β -D-xylopyranoside (▲)], conforming to Michaelis–Menten kinetics. The apparent Michaelis parameters (Table 3) were calculated based directly on the Michaelis–Menten equation.

Biochemical characterization of mature Bgl1 protein

The kinetic parameters K_m and V_{max} were determined for the mature protein (expressed from pBgl1_p in *E. coli* SURE cells) with various nitrophenyl (NP) glycosides, natural aryl glucosides arbutin and salicin, and various saccharides (Table 3). The highest catalytic efficiency, based on the k_{cat}/K_m ratio, was observed with the nitrophenyl glucosides, in particular for pNPG, with the catalytic efficiency for the natural aryl- β -glucosides, arbutin and salicin, 5.47- and 43.4-fold lower respectively. The enzyme showed no detectable activity toward substrates derived from cellulose hydrolysis, lactose and pNP β -D-glucuronide, pNP β -D-mannopyranoside and pNP α -L-rhamnopyranoside (Table 3, Figure 3). There was a clear preference for the hydrolysis of substrates with a β -linked glucose at the non-reducing end, and significant activities were observed for substrates with a β -linked galactose or xylose, but not mannose. Taking the aglycone side into account, there was no activity for any of the disaccharides tested, the natural compound arbutin being the preferred substrate.

Figure 4 Effect of temperature (A) and pH (B) on the activity and stability of the purified mature fused His₆-Bgl1 protein towards pNPG

The effects on activity (○) and stability (●) are shown. In (A), a further independent assay of the effect of temperature on activity is shown (▲).

Under the enzyme assay conditions used (with up to 12 min incubation periods), the optimal temperature for maximal β -glucosidase activity at pH 6.2 was 50 °C (Figure 4A). However, at this high temperature the stability was very poor. At 40 °C, Bgl1 appeared to be fairly stable and the enzyme activity only suffered a 40 % reduction of the maximal activity (Figure 4A). The enzyme was found to retain full activity after storage at 4 °C for several months. The optimum pH at 37 °C was between 5.1 and 5.7, the enzyme being fairly stable within the pH range from 4.7 to 7.3 (Figure 4B).

Sequence alignment

From the analysis of the complete manual alignment, it is possible to group sequences based on similarities in their patterns of deletions and insertions. Three subgroups are clearly identified in the alignment of Figure 5: a eukarya/plant group (sequence from *Zea mays*, *Trifolium repens* and *Sinapis alba*), an archaea group (sequence from *Thermosphaera aggregans* and *Sulfolobus*

solfataricus) and a bacteria group (all the remaining sequences). The latter can be subdivided in two minor subgroups: one with GH1 sequences from *S. paucimobilis* (Bgl1), *Novosphingobium aromaticovorans* (Narom_orf_145.18 and Narom_orf_61.1), *Streptomyces rochei* and *Deinococcus radiodurans*, and the other with β -glucosidases from *Erwinia herbicola*, *Agrobacterium* sp., *Bacillus polymyxa* and the two 6-phospho- β -glucosidases from *Lactobacillus lactis* and *E. coli*.

All GH1 enzymes possess a $(\alpha/\beta)_8$ -barrel architecture, but there are small differences in the length of the interconnecting loops (Figure 5). Interestingly, the sequences close to *S. paucimobilis* Bgl1 tend to have shorter loops interconnecting the different structural elements. The loops and β -strands that contain residues participating in subsite -1, found at the centre of the barrel, have more conserved sequence and are less prone to include insertions or deletions.

Active site signatures and subsites

The regions surrounding the active site, in particular those constituting the binding and catalytic sites, are well conserved, but contain a few differences that have an impact on enzyme and protein functionality. The region of the catalytic acid, Asn²⁵³–Glu²⁵⁴ in the alignment (Figure 5), is well conserved among all GH1 proteins, except in plant myrosinases and mammalian klotho proteins (the activities of which are still unknown), where the pair Asn-Glu is replaced by Asn-Gln and Asp-Asn respectively (results not shown). In myrosinases, the hydrolytic activity may be 'rescued' by an ascorbate placed in a critical position in the active site [11].

The more conserved regions in the active site can be clearly identified in the alignment (Figure 5). The region containing the catalytic nucleophile corresponding to [LIVMST]⁵⁰⁴ENG⁵⁰⁷ (alignment numbering; residues in brackets signify the possible amino acids at position 504), as partially described by PROSITE motif PS00572, is well conserved among GH1 members. However, several cases deviate from this 'consensus'. In particular, *S. paucimobilis* Bgl1 has a His residue immediately after the catalytic nucleophile Ser⁵⁰⁴-Glu-His-Gly⁵⁰⁷, a feature conserved only in Bgl1 and in the two *N. aromaticovorans* enzymes, but not seen in other GH1 enzymes. Significantly, different residues are found at other positions of this 'consensus' in the divergent *D. radiodurans* putative β -glucosidase, which contains the sequence Ala-Glu-Thr-Gly.

From the structural analysis of complexes between *B. polymyxa* (PDB structure 1E4I), *Z. mays* [13,14] and *S. alba* [11] GH1 enzymes and different substrates and reaction intermediates, it was possible to identify the active-site residues that form the different subsites according to the accepted nomenclature [12]. Subsite -1, or glycone subsite, is delimited in all three cases, but the subsite +1, or aglycone-binding subsite, could only be described for the *Z. mays* enzyme. The residues of both subsites are boxed in Figure 5. In this alignment, it is possible to observe most aspects related to substrate conservation and specificity among GH1 enzymes. In general, the residues which accommodate the non-reducing end saccharide moiety, and those which constitute subsite -1, tend to be very conserved. An exception to this conservation is found among 6-phospho- β -glucosidases, in which the Glu³⁶³ (alignment numbering) is replaced by a Ser residue to accommodate the more bulky 6-phosphate group [40], a feature found in all known phospho-carbohydrate-degrading enzymes present in this family (results not shown). The reason why only the *D. radiodurans* putative enzyme has a Gly residue there remains unclear, but this protein has so many different features that it is difficult to give attention

to specific details. For the most common β -glycosidase activities, no specific features associated with preferred monosaccharide moieties have arisen from the analysis of the information in Figure 5. Very recently, two specific amino acids related to β -mannosidase activity were clearly identified, Gln¹²⁰ and Asp²⁵³ (alignment numbering) instead of Arg¹²⁰ and Asn²⁵³, which improves the interaction with the axial hydroxyl group 2-OH of mannosyl residues at subsite -1 [41]. Differences at these positions are likely to affect the reaction steps expected for different glycosides in catalysis [41,42].

Apart from the deviations to the consensus around the catalytic acid already mentioned, the -1, or glycone, subsite is highly conserved among all GH1 proteins with β -glucosidase activity. For the +1, or aglycone, subsite, the panorama changes dramatically, as no significant set of residues is conserved, as seen in Figure 5. These results are consistent with the fact that GH1 enzymes have a large range of potential substrates and specificities, particularly in the nature of the aglycone. Only one residue appears more frequently in the sequence of the different GH1 enzymes, although not in *S. paucimobilis* Bgl1, and this is the Trp residue at positions 477 and 478 of the alignment (Trp⁴³² and Trp³²⁶ in *Z. mays* and in *B. polymyxa* β -glucosidases, respectively), which occurs at the interface of subsites -1 and +1, and is likely to constitute a stacking surface for sugar and aromatic residues binding at this subsite. Analysis of the region containing this Trp residue in the different known structures revealed a large variation among GH1 enzymes (results not shown).

Phylogenetic analysis

The mid-point rooted tree obtained in the phylogenetic analysis (Figure 6) reveals that the clustering of GH1 enzymes is in agreement with the phylogenetic group of the organism to which they belong. In grouping all plant enzymes, the major GH1 subfamily was named E1 (Eukarya 1) and this subfamily is characterized by different β -glycosidase and myrosinase activities. The remaining eukaryote GH1 enzymes share the same segment of the tree, consisting of five other subfamilies. These include enzymes from fungi (subfamily E6), insects (subfamily E3) and mammals (subfamilies E2, E4 and E5) (Figure 6). Among the mammalian proteins, there are some very unique activities, such as lactase phlorizin hydrolases and protein klotho [43].

The GH1s from archaea and bacteria are subdivided into two and six distinct subfamilies, referred to as A1–A2 and B1–B6 respectively (Figure 6). The major archaeal GH1 subfamily was designated A1 and is composed of enzymes with β -glucosidase, β -galactosidase and β -mannosidase activities. In contrast, the smaller A2 subfamily is exclusively composed of β -glucosidases. Bacterial GH1 enzymes with β -glucosidase activity are found in subfamilies B1, B2 and B6; the 6-phospho- β -glucosidases are preferentially clustered in subfamily B4, although some are present in subfamily B5. The bacterial subfamily B3 is exclusively composed of enzymes whose function has yet to be established.

Bgl1 from *S. paucimobilis* is found in the bacterial subfamily B6, together with putative β -glucosidases from *N. aromaticovorans* and a poorly characterized β -glucosidase from *S. rochei* [44]. This group is remotely related to all the other GH1 enzymes. Interestingly, subfamily B6 enzymes are more closely related to the subfamilies A1 and A2 than to the other bacterial subfamilies. Except for Bgl1 from *S. paucimobilis*, no other members of the B6 subfamily have been extensively characterized by biochemical methods.

Figure 5 Structure-based alignment of a representative set of GH1 sequences (For legend, see facing page.)

DISCUSSION

Enzyme activity and selectivity

In the present study, the *S. paucimobilis* β -glucosidase Bgl1 was examined as a representative member of a new GH1 subfamily. The existence of an N-terminal signal peptide of 24 amino acids, rich in hydrophobic residues, was investigated. The results strongly suggest that there the predicted signal does exist, that the protein undergoes processing and that the mature Bgl1 in *S.*

paucimobilis is located in the periplasm. Signal peptides are uncommon among bacterial and archaeal GH1 enzymes, suggesting that most are cytoplasmic. However, both Bgl1 and the two *Novosphingobium* putative β -glucosidases discussed here appear to be periplasmic. Interestingly, our results suggest that the loss of the signal peptide could be essential for enzyme activity in a singular activation process.

In the absence of structural data for complexes with substrates or their analogues covering the aglycone subsite, predictions of

Figure 5 Continued

In this alignment, the emphasis is on sequences closely related to *S. paucimobilis* Bgl1, and it includes all significant structures found at present. For available three-dimensional structures, the secondary structure is described by STRIDE [29], and is indicated in the Figure using background shading with the average secondary structure elements graphically represented at the bottom: α -helices or 3_1 -helices, light grey/cylinders; β -strands or β -bridges, dark grey/arrows. The glycone, or -1, subsite is represented by full-length vertical boxes and the conserved residues are indicated in the 'Active_site' row of the alignment. The aglycone, or +1, subsite is simply inferred by distance from the complex described by the Zmays_bgl_1E1EA (PDB:1E1E chain A) structure, the corresponding residues being individually boxed in the sequence (and in other sequences when identical), and an asterisk identifies the active site. The residues in contact, at a distance of less than 4.0 Å as determined by LPC software [30], with substrates or their analogues in the PDB structures (1E56, 1E41 and 1E73) are shown in bold. Sequences are identified by the following criteria: species are represented by the first letter of the genus followed by the first 4 letters of the remainder of the species name; this is followed by a three-letter abbreviation of the catalytic activity (see Figure 6 for a full list of activities and their abbreviations), and completed by a representative identification code issued from the PDB, SwissProt/TrEMBL or GenBank/EMBL protein identification (these identifiers are enough to identify the protein in the page corresponding to family GH1 in the CAZy server: <http://afmb.cnrs-mrs.fr/CAZY>). As an example, the β -glucosidase from *S. rochei*, found under the accession number Q55000 in TrEMBL is here identified as Sroch_bgl_Q55000. The *Novosphingobium aromaticivorans* sequences are simply labelled as Narom_orf_contig.gene. The N-terminal segments are shown only for the GH1 enzymes of species from *Sphingomonas* and *Novosphingobium* genera (*Sphingomonadaceae* family), with the putative signal peptides underlined.

Figure 6 Phylogenetic tree of the 286 GH1 sequences deposited in CAZy (January 2002), performed with TREECON [35], using the Neighbour-Joining method [36] supported by 500 bootstrap steps [37]

The phylogenetic tree was constructed based on the result of the global alignment of the 286 sequences performed by ClustalW 1.82. The mid-point root, automatically assigned by TreeCon, is represented by a dot. A branch length of 0.1 substitutions/site is given to infer phylogenetic distances. The individual sequences are identified as described in Figure 5. The GH1 proteins with structures already known are shown in bold. The identification of subfamilies following bootstrap analysis used *Athai_orf_AAG51335* as an out group. All sequences present in groups of 3 or more were assigned to a subfamily. Subfamilies were named according to kingdom (A, archaea; B, bacteria; E, eukarya) and by decreasing order in the number of sequences grouped. Sequences not classified in families have their identifiers in italics. Within each subfamily, groups of activities are also depicted. To avoid cluttering, only the sequences exhibiting less than 60% identity from others have their identity shown, except for those present in the manual alignment. The known activities are referenced as: bga, β -galactosidase (EC 3.2.1.23); bma, β -mannosidase (EC 3.2.1.25); bgl, β -glucosidase (EC 3.2.1.21); car, cardenolide 16-O-glucosylhydrolase; cyt, cytosolic β -glucosidase; dal, dalcocinin (β -glucosidase/ β -fucosidase); klo, klotho protein; mal, male-specific- β -glucosidase; lac, lactase-phlorizin hydrolase (EC 3.2.1.62); cya, cyanogenic β -glucosidase (EC 3.2.1.21); 6pb, 6-phospho- β -glucosidase (EC 3.2.1.86); pho, phospho- β -galactosidase (EC 3.2.1.85)/phospho- β -cellobiase (EC 3.2.1.-); prn, prunasin β -glucosidase (EC 3.2.1.118); myr, myrosinase (EC 3.2.3.1); rau, raucassicine-O-phospho- β -glucosidase (EC 3.2.1.125); sil, silverleaf whitefly-induced protein; str, strictosidine phospho- β -glucosidase (EC 3.2.1.105).

aglycone-specificity based on sequence similarities remain elusive. This is particularly true for Bgl1 and related proteins, which lack residues in common with those presently known to characterize this subsite. Even though Bgl1 is specific for aryl- β -glucosides, it shows no activity against cello-derivatives (and other sac-

charides), which contrasts with the traditional reference for microbial β -glucosidase activity, i.e. the degradation of cellobiose and other oligosaccharides derived from cellulose hydrolysis [4]. There are a growing number of GH1 enzymes that exhibit higher specificity towards aryl- β -glucosides, such as arbutin and salicin

[45], or their derivatives, in which the aglycone group is aromatic. Phenolic β -glucosides are among the compounds found in the leaves, bark and flowers of many plant species [46], and the presence of specific enzymic systems against such substrates is likely to reveal a specific interaction of the corresponding micro-organism with plants. Unsurprisingly, the *S. paucimobilis* strain examined in the present work was isolated from plant tissues [15]. Inspection of the chromosomal region surrounding the *bgl1* gene revealed that it is located downstream of the *ces10* gene, encoding an esterase [46a]. In bacteria and fungi that exploit plants as nutrient resources, esterases are known to remove a variety of attached groups from cell wall polysaccharides, facilitating the task of polysaccharide-degrading enzymes [47]. A similar co-ordinate action of Ces10 and Bgl1 proteins in the degradation of complex polysaccharides, namely those present in the aerial parts of plants colonized by *Sphingomonas* species [17], remains to be demonstrated.

Structure/function relationships

In the structure-based alignment, *S. paucimobilis* Bgl1 is clearly found in a subgroup distinct from the major archaeal/bacterial and eukaryotic groups. However, the alignment shows that structural features and specific residues involved in substrate binding and catalysis [48,49] are conserved. By homology, Bgl1 contains the structural elements characteristic of the $(\alpha/\beta)_8$ -barrel present in GH1 structures, some differences being observed in the loops interconnecting the different structural elements. It also conserves the main active-site characteristics of the GH1 family, in particular, the two catalytic glutamates (Glu²⁵⁴ and Glu⁵⁰⁵, alignment numbering) involved in the hydrolysis reaction. Interestingly, the catalytic nucleophile region exhibits a unique feature. Usually described by the conserved segment Thr⁵⁰⁴-Glu-Asn-Gly⁵⁰⁷, here it is replaced by Ser-Glu-His-Gly, a feature only shared with the two *Novosphingobium* enzymes. In *Agrobacterium* sp. β -glucosidase, the replacement of the residue Asn by Ser in this consensus reduced the *p*NPG activity 3-fold [50], an indication that changes in this region have a significant effect. Since none of the β -glucosidases of *Novosphingobium* have been characterized, the importance of the residues surrounding the catalytic nucleophile is still unknown. Structural studies would help to elucidate the Bgl1 structure/function relationships.

Phylogenetic relationships

The use of the most internal branches defined with robustness values above 85% [37] strongly supports the definition of subfamilies and, since it is more objective than previous analysis [43], it could constitute a homogeneous system for defining subfamilies in large family systems, such as CAZy [51]. The integration of such a system, coupled whenever possible with established subfamilies, would enlarge the advantages of the CAZy family classification. As enzymes with the same activity tend to group together, such a family subdivision scheme may, within limits, reduce the range of activities to be considered for the still uncharacterized GH1 enzymes.

In the tree resulting from the phylogenetic analysis, the GH1 enzymes are clustered in agreement with the phylogenetic group of the corresponding organisms. This suggests that, unlike the findings for other GH families [52], no significant horizontal transfer of GH1 enzymes has occurred between phyla during evolution. In particular, the GH1 enzymes of bacterial and archaeal origin could be less attractive candidates for horizontal transfer: most are singular domain and intracellular (the few periplasmic cases, such as *S. paucimobilis* Bgl1, are the exception)

and many appear now to deal with intracellular substrates in specific metabolic and/or signalling roles.

With the exception of a more divergent *A. thaliana* ORF, used as a root to define the most internal branches, all other plant GH1 proteins are found in the E1 subfamily. This very large group has the largest diversity of enzyme activities among GH1 subfamilies, all of which are likely to have evolved for very specialized tasks. The full extent of this variety in *A. thaliana* and that of the complementary, and even more numerous, glycosyl-transferase family 1 (or GT1) has been recently catalogued [53]. A few particular enzyme activities, exclusive to animals, are found for other eukaryote subfamilies, in particular E2 and E5. The remaining eukaryote subfamilies, insect and fungal subfamilies E3 and E6 respectively, have more traditional β -glucosidase activities. Among the bacterial subfamilies B1–B5, enzymes with the same activity tend to group together. Subfamilies B4 and B5 deal exclusively with 6-phospho-carbohydrate substrates, and this is almost true for the B1 subfamily, which also contains a few β -glucosidases. Interestingly, no enzymes from subfamily B3 have been yet characterized, but the presence of the characteristic Ser residue that allows the accommodation of the phosphate group at subsite –1, referred to above, suggests that it is also composed of phospho-carbohydrate-degrading enzymes. It is now clear that subfamily B6, of which Bgl1 is a member, is a very divergent branch of GH1 enzymes.

This work was supported by Fundação para a Ciência e a Tecnologia (FCT), Portugal FEDER and PRAXIS XXI Program and Programa Operacional de Quadro Comunitário de Apoio (POCTI). In particular, we acknowledge projects POCTI/35733/1999 and POCTI/BCI/33854/1999 and grants to A.R.M. (PRAXIS XXI/BD/18146/98) and P.V. (SFRH/BPD/5710/2001). The contribution of Tiago Fernandes in the preparation of Figure 5 is acknowledged.

REFERENCES

- 1 Sinnott, M. L. (1990) Catalytic mechanisms of glycosyl transfer. *Chem. Rev.* **90**, 1171–1202
- 2 Zechel, D. and Withers, S. J. (2000) Glycosidase mechanisms: anatomy of a finely tuned catalyst. *Acc. Chem. Res.* **33**, 11–18
- 3 Faure, D., Henrissat, B., Ptacek, D., Bekri, M. A. and Vanderleyden, J. (2001) The *celA* gene, encoding a glycosyl hydrolase family 3 β -glucosidase in *Azospirillum irakense*, is required for optimal growth on cellobiosides. *Appl. Environ. Microbiol.* **67**, 2380–2383
- 4 Beguin, P. (1990) Molecular biology of cellulose degradation. *Annu. Rev. Microbiol.* **44**, 219–248
- 5 Hughes, M. A., Brown, K., Pancoro, A., Murray, B. S., Oxtoby, E. and Hughes, J. (1992) A molecular and biochemical analysis of the structure of the cyanogenic β -glucosidase (linamarase) from cassava (*Manihot esculenta* Cranz). *Arch. Biochem. Biophys.* **295**, 273–279
- 6 Henrissat, B. (1991) Classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J.* **280**, 309–316
- 7 Rojas, A., Arola, L. and Romeu, A. (1995) β -Glucosidase families revealed by computer analysis of protein sequences. *Biochem. Mol. Biol. Int.* **35**, 1223–1231
- 8 Coutinho, P. and 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* (Ohmura, K., Hayashi, K., Sakka, K., Kobayashi, Y., Karita, S. and Kimura, T., eds.), pp. 15–23, Uni Publishers Co., Tokyo
- 9 Henrissat, B., Callebaut, I., Fabrega, S., Lehn, P., Mornon, J. P. and Davies, G. (1995) Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 7090–7094
- 10 Reference deleted
- 11 Burmeister, W. P., Cottaz, S., Rollin, P., Vasella, A. and Henrissat, B. (2000) High resolution X-ray crystallography shows that ascorbate is a cofactor for myrosinase and substitutes for the function of the catalytic base. *J. Biol. Chem.* **275**, 39385–39393
- 12 Davies, G. J., Wilson, K. S. and Henrissat, B. (1997) Nomenclature for sugar-binding subsites in glycosyl hydrolases. *Biochem. J.* **321**, 557–559

- 13 Czjzek, M., Cicek, M., Zamboni, V., Bevan, D. R., Henrissat, B. and Esen, A. (2000) The mechanism of substrate (aglycone) specificity in β -glucosidases is revealed by crystal structures of mutant maize β -glucosidase-DIMBOA, -DIMBOAGlc, and -dhurrin complexes. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 13555–13560.
- 14 Czjzek, M., Cicek, M., Zamboni, V., Burmeister, W. P., Bevan, D. R., Henrissat, B. and Esen, A. (2001) Crystal structure of a monocotyledon (maize ZMGlut1) β -glucosidase and a model of its complex with *p*-nitrophenyl β -D-thioglucoiside. *Biochem. J.* **354**, 37–46.
- 15 Kang, K. S., Veeder, G. T., Mirrasoul, P. J., Kaneko, T. and Cottrell, W. (1982) Agar-like polysaccharide produced by a *Pseudomonas* species: production and basic properties. *Appl. Environ. Microbiol.* **43**, 1086–1091.
- 16 Takeuchi, M., Sakane, T., Yanagi, M., Yamasato, K., Hamana, K. and Yokota, A. (1995) Taxonomic study of bacteria isolated from plants: proposal of *Sphingomonas rosa* sp. nov., *Sphingomonas pruni* sp. nov., and *Sphingomonas mali* sp. nov. *Int. J. Syst. Bacteriol.* **45**, 334–341.
- 17 Kim, H., Nishiyama, M., Kunito, T., Senoo, K., Kawahara, K., Murakami, K. and Oyaizu, H. (1998) High population of *Sphingomonas* species on plant surface. *J. Appl. Microb.* **85**, 731–736.
- 18 Yun, N. R., Shin, Y. K., Hwang, S. Y., Kuraishi, H., Sugiyama, J. and Kawahara, K. (2000) Chemotaxonomic and phylogenetic analyses of *Sphingomonas* strains isolated from ears of plants in the family *Gramineae* and a proposal of *Sphingomonas roseoflava* sp. nov. *J. Gen. Appl. Microbiol.* **46**, 9–18.
- 19 Videira, P. A., Cortes, L. L., Fialho, A. M. and Sá-Correia, I. (2000) Identification of *pgmG* gene, encoding a bifunctional protein with phosphoglucomutase and phosphomannomutase activities, in gellan gum producing strain *Sphingomonas paucimobilis* ATCC 31461. *Appl. Environ. Microbiol.* **66**, 2252–2258.
- 20 Videira, P. A., Fialho, A. M., Geremia, R. A., Breton, C. and Sá-Correia, I. (2001) Biochemical characterization of the β -1,4-glucuronosyltransferase GelK in gellan-gum producing strain *Sphingomonas paucimobilis* A.T.C.C. 31461. *Biochem. J.* **358**, 457–464.
- 21 Sá-Correia, I., Fialho, A. M., Videira, P., Moreira, L. M., Marques, A. R. and Albano, H. (2002) Gellan gum biosynthesis in *Sphingomonas paucimobilis* ATCC 31461: genes, enzymes and exopolysaccharide production engineering. *J. Ind. Microbiol. Biotechnol.* **29**, 170–176.
- 22 Birnboim, H. C. and Doly, J. (1979) A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Res.* **7**, 1513–1523.
- 23 Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) *Molecular Cloning: a Laboratory Manual*, 2nd edn, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
- 24 Schirmer, F., Ehrst, S. and Hillen, W. (1997) Expression, inducer spectrum, domain structure and function of MopR, the regulator of phenol degradation in *Acinetobacter calcoaceticus*. *J. Bacteriol.* **179**, 1329–1336.
- 25 French, C., Keshavarz-Moore, E. and Ward, J. (1996) Development of a simple method for the recovery of recombinant proteins from the *Escherichia coli* periplasm. *Enzyme Microb. Technol.* **19**, 332–338.
- 26 Bradford, M. M. (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72**, 248–254.
- 27 Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
- 28 Kyte, J. and Doolittle, R. F. (1982) A Simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–142.
- 29 Frishman, P. and Argos, P. (1995) Knowledge-based protein secondary structure assignment. *Proteins* **23**, 566–579.
- 30 Sobolev, V., Sorokina, A., Prilusky, J., Abola, E. E. and Edelman, M. (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatics* **15**, 327–332.
- 31 Mizuguchi, K., Deane, C. M., Blundell, T. L. and Overington, J. P. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.* **7**, 2469–2471.
- 32 De Rijk, P. and Wachter, R. De (1993) DCSE, an interactive tool for sequence and secondary structure research. *Comput. Appl. Biosci.* **9**, 735–740.
- 33 Lemesle-Verloot, L., Henrissat, B., Garboraud, C., Bissery, V., Morgat, A. and Mornon, J.-P. (1990) Procedures to derive structural and functional information from 2-D representation of protein sequences. *Biochimie* **72**, 555–574.
- 34 Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.
- 35 Van de Peer, Y. and De Wachter, R. (1994) TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. *CABIOS Comput. Appl. Biosci.* **10**, 569–570.
- 36 Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.
- 37 Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791.
- 38 Yamazaki, M., Thorne, L., Mikolajczak, M., Armentrout, R. W. and Pollock, T. J. (1996) Linkage of genes essential for synthesis of a polysaccharide capsule in *Sphingomonas* strain S88. *J. Bacteriol.* **178**, 2676–2687.
- 39 Peciña, A., Pascual, A. and Paneque, A. (1999) Cloning and expression of the *algL* gene, encoding the *Azotobacter chroococcum* alginate lyase: purification and characterization of the enzyme. *J. Bacteriol.* **181**, 1409–1414.
- 40 Sanz-Aparicio, J., Hermoso, J. A., Martínez-Ripoll, M., Lequerica, J. L. and Polaina, J. (1998) Crystal structure of β -glucosidase A from *Bacillus polymyxa*: insights into the catalytic activity in family 1 glycosyl hydrolases. *J. Mol. Biol.* **275**, 491–502.
- 41 Kaper, T., van Heusden, H. H., van Loo, B., Vasella, A., van der Oost, J. and de Vos, W. M. (2002) Substrate specificity engineering of β -mannosidase and β -glucosidase from *Pyrococcus* by exchange of unique active site residues. *Biochemistry* **41**, 4147–4155.
- 42 Ducros, V. M., Zechel, D. L., Murshudov, G. N., Gilbert, H. J., Szabo, L., Stoll, D., Withers, S. G. and Davies, G. J. (2002) Substrate distortion by a β -mannanase: snapshots of the Michaelis and covalent-intermediate complexes suggest a β (2,5) conformation for the transition state. *Angew. Chem. Int. Ed.* **41**, 2824–2827.
- 43 Mian, I. S. (1998) Sequence, structural, functional, and phylogenetic analyses of three glycosidase families. *Blood Cell Mol. Dis.* **24**, 83–100.
- 44 Mastromei, G., Hanhart, E., Perito, B. and Polsinelli, M. (1992) Cloning and expression in *Escherichia coli* of a *Streptomyces* β -glucosidase gene. *J. Biotech.* **24**, 149–157.
- 45 Marri, L., Valentini, S. and Venditti, D. (1995) Cloning and nucleotide sequence of the *bglA* gene from *Erwinia herbicola* and expression of β -glucosidase activity in *Escherichia coli*. *FEMS Microbiol. Lett.* **128**, 135–138.
- 46 Bender, C. L., Alarcon-Chaidez, F. and Gross, D. C. (1999) *Pseudomonas syringae* phytoalexins: mode of action, regulation, and biosynthesis by peptide and polyketide synthetases. *Microbiol. Mol. Biol. Rev.* **63**, 266–292.
- 46a Videira, P. A., Fialho, A. M., Marques, A. R., Coutinho, P. M. and Sá-Correia, I. (2003) Cloning and sequence analysis of the *ces10* gene encoding a *Sphingomonas paucimobilis* esterase. *Appl. Microbiol. Biotechnol.*, in the press.
- 47 Williamson, G., Kroon, P. A. and Faulds, C. B. (1998) Hairy plant polysaccharides: a close shave with microbial esterases. *Microbiology* **144**, 2011–2023.
- 48 Henrissat, B. and Bairoch, A. (1996) Updating the sequence-based classification of glycosyl hydrolases. *Biochem. J.* **316**, 695–696.
- 49 Henrissat, B. and Davies, G. (1997) Structural and sequence-based classification of glycosyl hydrolases. *Curr. Opin. Struct. Biol.* **7**, 637–644.
- 50 Trimbur, D. E., Warren, R. A. and Withers, S. G. (1992) Region-directed mutagenesis of residues surrounding the active site nucleophile in β -glucosidase from *Agrobacterium faecalis*. *J. Biol. Chem.* **267**, 10248–10251.
- 51 Henrissat, B. and Romeu, A. (1995) Families, superfamilies and subfamilies of glycosyl hydrolases. *Biochem. J.* **311**, 350–351.
- 52 García-Vallvé, S., Romeu, A. and Palau, J. (2000) Horizontal gene transfer of glycosyl hydrolases of the rumen fungi. *Genome Res.* **10**, 1719–1725.
- 53 Henrissat, B., Coutinho, P. M. and Davies, G. J. (2001) A census of carbohydrate-active enzymes in the genome of *Arabidopsis thaliana*. *Plant Mol. Biol.* **47**, 55–72.

Received 8 August 2002/13 November 2002; accepted 22 November 2002

Published as BJ Immediate Publication 22 November 2002, DOI 10.1042/BJ20021249