PROTEINS: Structure, Function, and Genetics 42:66–76 (2001)

β-Fructosidase Superfamily: Homology With Some α-L-Arabinases and β-D-Xylosidases

Daniil G. Naumoff*
State Institute for Genetics and Selection of Industrial Microorganisms, Moscow, Russia

ABSTRACT **Comparison of the amino acid sequences of four families of glycosyl hydrolases reveals that they are homologous and have several common conserved regions. Two of these families contain β-fructosidases (glycosyl hydrolase families GH32 and GH68) and the other two include α-L-arabinases and β-xylosidases (families GH43 and GH62). The latter two families are proposed to be grouped together with the former two into the β-fructosidase (furanosidase) superfamily. Several ORFs can be considered as a fifth family of the superfamily on the basis of sequence similarity. It is shown for the first time that a glycosyl hydrolase superfamily can include enzymes with both inversion and retention mechanism of action. Composition of the active center for enzymes of the superfamily is discussed. Proteins 2001;42:66–76.**

Key words: protein family; glycosyl hydrolase; furanosidase; levansucrase; sucrase; hydrophobic cluster analysis; multiple sequence alignment

INTRODUCTION

Glycoside hydrolases or glycosidases (EC 3.2.1.-) are a widespread group of enzymes of significant biochemical, medical, and industrial importance that hydrolyze the glycosidic bonds between two carbohydrates or between a carbohydrate and an aglycone moiety. A large multiplicity of these enzymes is a consequence of the extensive variety of their natural substrates: di-, oligo-, and polysaccharides. The traditional nomenclature of glycosidases[1] is based on their substrate specificity and occasionally on the molecular mechanism of their action; such a classification, however, does not reflect the structural features and evolutionary relationships of these enzymes, and it is not appropriate for enzymes that act on several substrates.[2–6]

Comparative analysis of 300 amino acid sequences of glycosidases known at the beginning of the 1990s showed that they could be classified into 36 families.[2] Progress in sequence data for glycosidases enables the discovery of new families. Currently, several thousand sequences of glycosidases and related proteins (transglycosidases, etc.) are known, and they have been grouped into 78 families.[3–11] Each of the families includes similar proteins over a fragment of >100 amino acid residues.[2] The basic principle underlying the classification is that the family membership can be established on the basis of a sequence alone.[2,7] Glycosidases catalyze the hydrolysis of the glycosidic bond of their substrates via two general mechanisms, leading to either inversion or overall retention of the anomeric configuration at the cleavage point.[4–6,11–14] With no known exceptions, the mechanism is conserved among all members of a given family.[3–6,12,13,15] The stereochemistry of hydrolysis reaction is known for at least 50 families.[5,6,9–12]

Detailed comparison of protein structures reveals some similarities between representatives of different families.[3–6,8–11,16–18] The related families are grouped at a higher hierarchical level into superfamilies (or clans). About a dozen of glycoside hydrolase superfamilies have been described. The largest of them (clan GH-A) includes 14 families; most others consist of 2 families each.[10,11]

Recently, we showed that β-fructosidases belonging to families GH32 (sucrases and related enzymes) and GH68 (levansucrases) are homologous.[19] It allowed us to combine them into the β-fructosidase superfamily (clan GH-J according to the classification of Coutinho and Henrissat[11]). Enzymes of both families have the same molecular mechanism of hydrolyzing reaction: double displacement with overall retention of the anomeric configuration of the β-D-fructofuranosyl residue,[10,11] and their sequences have nine common conserved regions.[19]

According to our preliminary data, bifunctional β-xylosidases and α-L-arabinofuranosidases of family GH43 and levansucrases are similar: they have two common sequence motifs.[20] It was recently noticed that glycosidases of family GH43 and α-L-arabinofuranosidases of family GH62 have some sequence similarity and compose a superfamily (clan GH-F).[11] In the present article, on the basis of detailed comparison of primary and hypothetical secondary structures, we show that sequences of glycoside hydrolases from families GH32, GH43, GH62, and GH68 have several common conserved regions and, therefore, compose a superfamily.

MATERIALS AND METHODS

Protein and nucleic sequences were retrieved from the current sequence databases. Proteins compared in this work are listed in Table I. Alignments of protein sequences were generated by using the PSI-BLAST program.[21] The

*Correspondence to: State Institute for Genetics and Selection of Industrial Microorganisms, I-Dorozhny proezd, 1, Moscow 113545, Russia. E-mail: daniil_naumoff@yahoo.com

Received 8 May 2000; Accepted 29 August 2000

TABLE I. Glycosyl Hydrolases Analyzed in the Work[a]

Family[b]	Organism	Enzyme[g]	Length[h]	Accession number[j]
43a	*Bacteroides ovatus*	α-L-arabinofuranosidase, Xylan 1,4-β-xylosidase	325	P49943
43a	*Cochliobolus carbonum*	Xylan 1,4-β-xylosidase	328	AAC67554*
43a	*Prevotella bryantii (ruminicola)* $B_1$4	Xylan 1,4-β-xylosidase	319	P48791
43a	*Clostridium stercorarium*	α-L-arabinofuranosidase, Xylan 1,4-β-xylosidase	473	P48790, JQ1936*
43b	*Pseudomonas fluorescens*	1,5-α-L-arabinosidase	347	P95470
43b	*Bacillus subtilis* 168	ORF	313	P94522
43b	*Bacillus subtilis* IFO3134	1,5-α-L-arabinosidase	324	O07078
43b	*Aspergillus niger*	1,5-α-L-arabinosidase	321	P42256
43b	*Bacillus subtilis* 168	ORF	469	P42293
43b	*Streptomyces coelicolor* A3(2)	ORF	322	CAB92901*
43c	*Bacillus subtilis* 168 and *Bacillus* sp. KK-1	Xylan 1,4-β-xylosidase	533	P94489, O52729
43c	*Bacillus pumilus* IPO and PLS	Xylan 1,4-β-xylosidase	535	P07129, AAC97375*
43c	*Selenomonas ruminantium*	α-L-arabinofuranosidase, Xylan 1,4-β-xylosidase	538	O52575
43c	*Escherichia coli*	ORF	536	P77713
43c	*Butyrivibrio fibrisolvens* GS113	α-L-arabinofuranosidase, Xylan 1,4-β-xylosidase	517	P45982
43c	*Streptomyces coelicolor* A3(2)	ORF	509	CAB52932*
43c	*Streptomyces coelicolor* A3(2)	ORF	95[i]	CAB46384*
43c	*Lactococcus lactis* 210, IO-1, and NRRL B4449	Xylan 1,4-β-xylosidase	269[i], 152[i], 257[i]	AAD20247*, AAD20253*, AAD20259*
43c	*Prevotella (Bacteroides) ruminicola* T31	Xylan 1,4-β-xylosidase	452	BAA78558*
43c	*Azospirillum irakense*	α-L-arabinofuranosidase, Xylan 1,4-β-xylosidase	542	AAF66622*
43c/d[c]	*Caldicellulosiruptor saccharolyticus (Caldocellum saccharolyticum)*	α-L-arabinofuranosidase, Xylan 1,4-β-xylosidase	1347	O30426
10/43d[d]	*Caldicellulosiruptor* sp. Rt69B.1 and Tok7B.1	α-L-arabinofuranosidase, Endo-1,4-β-xylanase	1779, 1770	O52374, Q9X3P5
43d	*Paenibacillus (Bacillus) polymyxa*	Endo-1,4-β-xylanase	635	P45796
43d	*Bacillus subtilis* 168	ORF	513	Q45071
43e	*Butyrivibrio fibrisolvens* H17c	α-L-arabinofuranosidase	789	Q45134
43e	*Ustilago maydis*	ORF	356	Q92388
43f	*Arabidopsis thaliana*	ORF	466, 239[i]	CAB66926*, CAA10760*
43[e]	*Streptomyces chartreusis*	α-L-arabinofuranosidase	328	BAA90772*
43[e]	*Salmonella typhimurium*	ORF	316	CAB89837*
43[e]	*Streptomyces coelicolor* A3(2)	ORF	370	CAB61805*
10/62[f]	*Streptomyces chattanoogensis*	α-L-arabinofuranosidase, Endo-1,4-β-xylanase	819	AAD32559*
62	*Streptomyces lividans* and *Streptomyces coelicolor* A3(2)	α-L-arabinofuranosidase	475	P96463, O54161
62	*Aspergillus sojae*	α-L-arabinofuranosidase	328	BAA85252*
62	*Aspergillus tubingensis* and *Aspergillus niger*	α-L-arabinofuranosidase	332	P79021, P79019
62	*Pseudomonas fluorescens*	α-L-arabinofuranosidase	571	P23031
GHLP	*Thermotoga maritima*	ORF	296	AAD36914*
GHLP	*Pyrococcus horikoshii*	ORF	299	BAA30206*
GHLP	*Pyrococcus abyssi*	ORF	305	CAB50037*
GHLP	*Thermotoga maritima*	ORF	326	AAD36300*
GHLP	*Thermotoga maritima*	ORF	334	AAD35864*
GHLP	*Aquifex aeolicus*	ORF	349	AAC07180*
GHLP	*Aeropyrum pernix*	ORF	366	BAA79294*
GHLP	*Mycobacterium tuberculosis*	ORF	299	P71783
32a	*Thermotoga maritima*	Invertase, Inulinase	432	O33833
32a	*Vibrio alginolyticus*	Invertase	484	P13394
32a	*Escherichia coli*	Invertase	477	O86076
32a	*Salmonella typhimurium*	Invertase	466	P37075
32a	*Pediococcus pentosaceus* and *Lactobacillus plantarum*	Invertase	501, 231[i]	P43471, O69442
32a	*Streptococcus mutans* GS-5	Invertase	454	P13522

TABLE I. (Continued)

Family[b]	Organism	Enzyme[g]	Length[h]	Accession number[j]
32a	*Bacillus subtilis* 168	Invertase	480	P07819
32a	*Leishmania major* Friedlin	ORF	513	CAB55619*
32b	*Bacillus subtilis* 168	Levanase	677	P05656
32b	*Streptococcus mutans* GS-5	Fructan β-fructosidase	1423	Q03174
32b	*Actinomyces naeslundii*	Levanase	943	Q44109
32b	*Bacteroides fragilis*	Levanase	622	Q45155
32b	*Arthrobacter nicotinovorans*	Levan fructosyltransferase	517	O50585
32b	*Tritrichomonas foetus*	Invertase	550	O02490
32b	*Saccharomyces cerevisiae*	Invertase	532	P00724
32b	*Kluyveromyces marxianus (fragilis)*	Inulinase	555	P28999
32b	*Debaryomyces (Schwanniomyces) occidentalis*	Invertase	533	P24133
32b	*Schizosaccharomyces pombe*	Invertase	581	O59852
32b	*Penicillium purporogenum*	Inulinase	515	O00056
32b	*Aspergillus foetidus*	Sucrose/sucrose 1-fructosyltransferase	537	O42801
32c	*Aspergillus niger*	Invertase	589	S33920*
32d	*Bacillus circulans*	Cycloinulo-oligosaccharide fructanotransferase	1503	O52973
32d	*Leishmania major* Friedlin	ORF	640	CAA21433*
32d	*Allium cepa*	Fructan/fructan 6G-fructosyltransferase	612	P92916
32d	*Helianthus tuberosus*	1,2-β-fructan 1F-fructosyltransferase	615	O81985
32d	*Daucus carota*	Invertase	592	P26792
32d	*Zea mays*	Invertase	597	AAD02264*
32d	*Chenopodium rubrum*	Invertase	573	Q42691
32d	*Vigna radiata (Phaseolus aureus)*	Invertase	649	P29001
32d	*Lycopersicon esculentum*	Invertase	582	O82119
68a	*Bacillus subtilis* 168 and *Bacillus stearothermophilus*	Levansucrase	473	P05655, P94468
68a	*Bacillus amyloliquefaciens*	Levansucrase	472	P21130
68a	*Bacillus* sp. V230	β-fructosyltransferase	487	O82854
68a	*Paenibacillus (Bacillus) polymyxa*	Levansucrase	499	CAB39327*
68a	*Streptococcus mutans*	β-fructosyltransferase	797	P11701
68a	*Streptococcus salivarius*	β-fructosyltransferase	969	Q55242
68b	*Gluconacetobacter (Acetobacter) diazotrophicus*	Levansucrase	584	Q43998
68b	*Erwinia amylovora* and *Rahnella aquatilis*	Levansucrase	415	Q46654, O54435
68b	*Pseudomonas syringae* pv. *glycinea*	Levansucrase	415	O52408
68b	*Pseudomonas syringae* pv. *phaseolicola*	Levansucrase	431	O68609
68b	*Acetobacter xylinus*	Levansucrase	430	BAA93720*
68b	*Zymomonas mobilis* NRRL B806, ATCC10988, and IFO13756	Levansucrase	423	Q60114, S33771*, JC2519*
68b	*Zymomonas mobilis* NRRL B806, ATCC10988, and IFO13756	Invertase	413	AAC36942*, S47527*, Q60115, BAA04476*

[a]All known sequences of families GH43, GH62, GH68, and GHLP and divergent representatives of family GH32, which includes about 200 sequences, are listed.[11]

[b]Family enumeration is given according to the classification of glycosyl hydrolases.[10,11] Classification for subfamilies of family GH43 (a–f) is according to the data of the present study. Sequences of 43a subfamily were proposed to consider as a separate family.[52,53] Classification for subfamilies of families GH32 (a–d) and GH68 (a, b) is according to Pons et al.[36] and our unpublished data. GHLP family is a glycosyl hydrolase like protein family described in the text.

[c]The protein consists of two homologous domains (see Figs. 1, 2, and 4).

[d]The C-terminal domain belongs to family GH43 and the N-terminal domain belongs to family GH10.[10,11]

[e]This sequence has unusual structure for family GH43.

[f]The C-terminal domain belongs to family GH62 and the N-terminal domain belongs to family GH10.[11]

[g]Enzyme activities: Levansucrase (EC 2.4.1.10); Sucrose/sucrose 1-fructosyltransferase (EC 2.4.1.99); Fructan/fructan 1-fructosyltransferase (EC 2.4.1.100); Uncharacterized β-fructosyltransferase (EC 2.4.1.x); Inulinase (EC 3.2.1.7); Endo-1,4-β-xylanase (EC 3.2.1.8); Invertase (β-fructofuranosidase, EC 3.2.1.26); Xylan 1,4-β-xylosidase (exo) (EC 3.2.1.37); α-L-arabinofuranosidase (1,2, 1,3, and/or 1,5) (EC 3.2.1.55); Levanase (EC 3.2.1.65); Fructan β-fructosidase (EC 3.2.1.80); Arabinan endo-1,5-α-L-arabinosidase (EC 3.2.1.99).

[h]Number of amino acid residues in the preprotein.

[i]A partial sequence.

[j]Accession numbers indicated by asterisks are from GenPept; the others are from SwissProt.

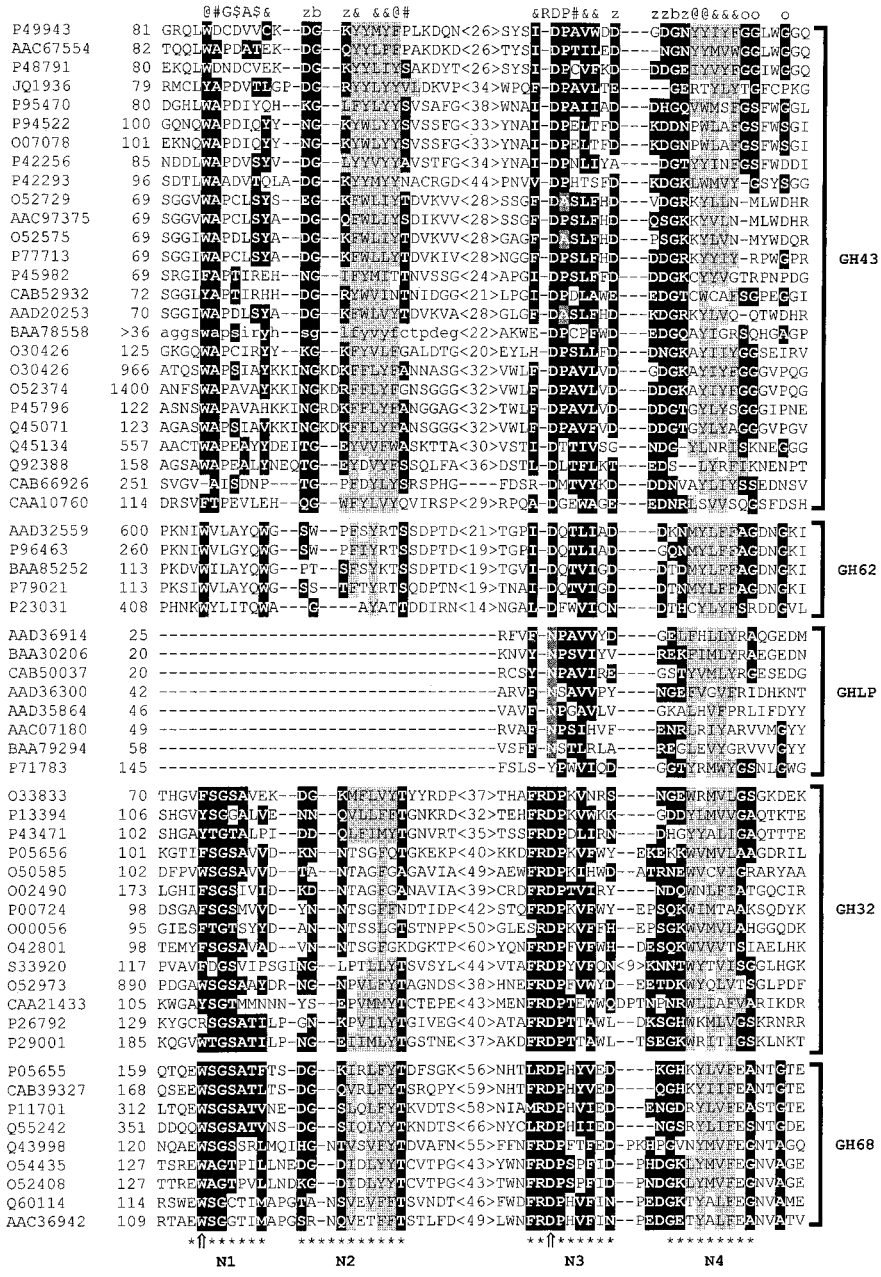

Fig. 1. Multiple sequence alignment of the sequences analyzed (fragment 2). Modified fragment of the sequence (BAA78558) obtained by improvement in frame shift alteration (see text) is shown by small letters. Clusters of bulky hydrophobic residues (I, L, V, M, F, Y, and W) are shaded (see text). At the bottom of the figure, conserved regions (N1–N4) are indicated by asterisks, and conserved residues discussed in the text are indicated by arrows. Other designations are the same as for Figure 4.

statistical significance threshold for including a sequence in the model (E-value) used by PSI-BLAST on the next iteration was 10^{-3}. Different alignments were compared manually. The MACAW program[22] was used to find regions of local similarity in different sequences. Regions of different sequences were considered similar if the probability of obtaining the observed level of similarity by chance (P value) was 10^{-4} or below. Hydrophobic cluster analysis[23] was used for predicting the secondary structure elements in conserved regions of sequences. The results obtained are consistent with the ones produced by the PredictProtein server (http://www.embl-heidelberg.de/predictprotein/predictprotein.html).[24]

RESULTS

On the basis of the preliminary data[11,20] about the similarities of glycosidases from family GH43 with some other enzymes, we compared sequences of this family with the whole current protein database. PSI-BLAST searches with a few randomly selected divergent representatives of

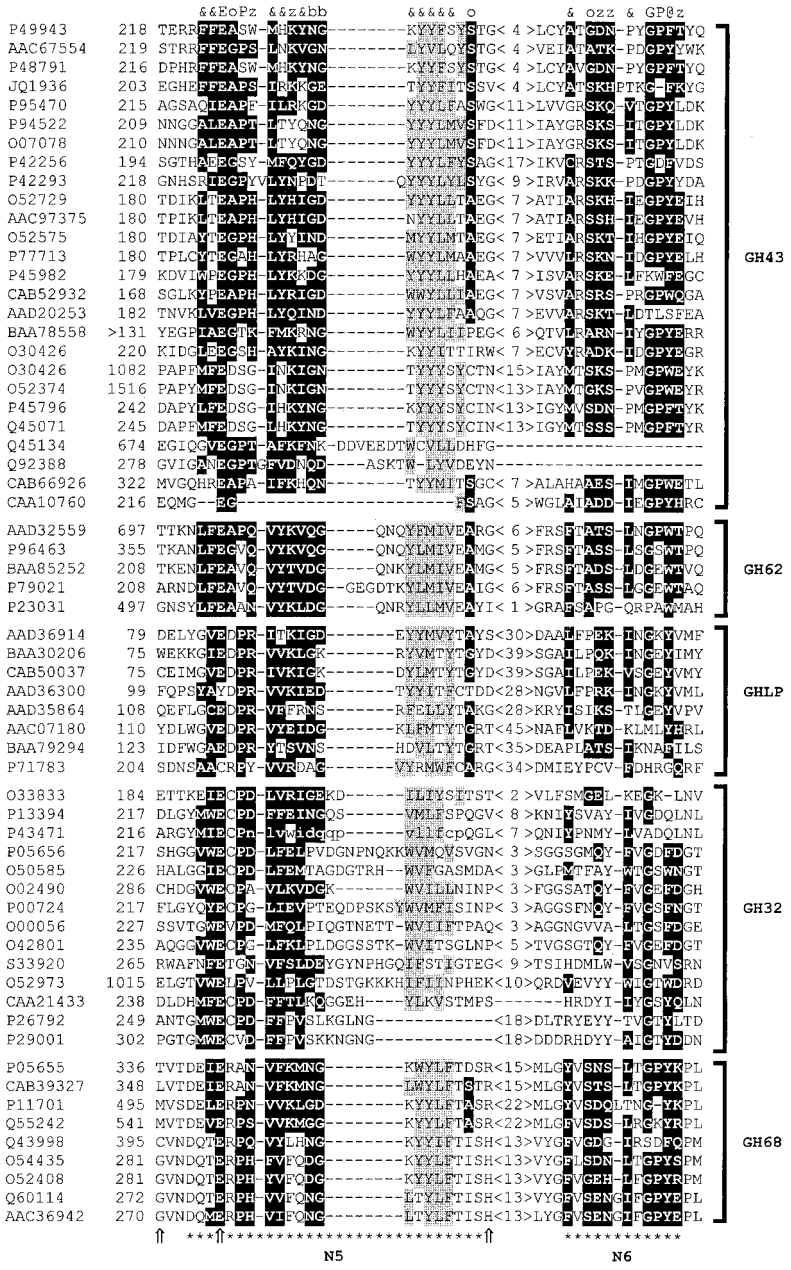

Fig. 2. Multiple sequence alignment of the sequences analyzed (fragment 3). Modified fragment of the sequence (P43471) obtained by improvement in frame shift alteration (see text) is shown by small letters. At the bottom of the figure, conserved regions (N5 and N6) are indicated by asterisks, and residues discussed in the text are indicated by arrows: catalytic conserved Glu, Cys of *Gluconacetobacter diazotrophicus* levansucrase (Q43998) involved in the disulfide bridge, and Arg/His of levansucrases essential for a polymerase activity. Other designations are the same as for Figures 1 and 4.

family GH43, as a query sequence, during the first round, and two iterations always yielded the complete set of proteins of the family with the exception of two sequences (SwissProt accession numbers Q45134 and Q92388) that probably were included[11] into the family on a basis of a low similarity level. In several cases during the first two iterations a hypothetical protein from *Arabidopsis thaliana* (GenPept accession numbers CAB66926 and CAA10760) was also yielded. Analysis of the order of the sequence appearances during searchers by PSI-BLAST, depending on the query, allows us to distinguish in family GH43 four subfamilies (43a–d) with at least five known members in each of them (Table I). The two divergent sequences compose the fifth subfamily (43e) and the *A. thaliana* hypothetical protein can be considered as the only representative of a sixth subfamily (43f).

Further PSI-BLAST iterations revealed weaker but still statistically significant similarities between members of family GH43 and glycosidases of families GH32, GH62, and GH68. In addition, some similarities with a number of

uncharacterized hypothetical proteins were found. PSI-BLAST searches using each of them as a query sequence showed that they form a group of homologous proteins, except an ORF from *Synechocystis* sp. (GenPept accession number BAA17165 is not considered in the present study). We called this group GHLP (glycoside hydrolase like protein) family.

Results of multiple sequence alignment of members of each family by PSI-BLAST revealed that some sequences had regions of local dissimilarities with the other sequences. Examination of the corresponding sites of nucleic sequences allowed us to improve the similarity by insertion or deletion of a single nucleotide. Several such frame shifts were described earlier in the sequences of family GH32: *Bacillus subtilis* sucrase (SwissProt accession number P07819), *B. stearothermophilus* levanase (P94469), *Bacillus* sp. L7 levanase (O31411), *Debaryomyces occidentalis* invertase (P24133), *Aspergillus niger* invertase (O13388), and *Avena sativa* invertase (Q43076).[25–30] To our knowledge, in two sequences frame shifts are found for the first time. One of them is the sequence of *Prevotella ruminicola* β-xylosidase (BAA78558), which has at least one frame shift in its N-terminal part that leads to missing of Trp-Ala amino acid pair, highly conserved in glycosidases of family GH43 (Fig. 1). Another is the sequence of *Pediococcus pentosaceus* sucrase (P43471). Its gene has 98.6% homology with *Lactobacillus plantarum* sucrase gene;[31] however, these two sucrases do not have homology in a 15-amino acid fragment. The sequence of this fragment of *L. plantarum* but not *P. pentosaceus* is similar to sucrase sequences from other bacteria that allowed us to improve the *P. pentosaceus* sucrase sequence (Fig. 2).

To examine a level of similarities among the five families, we compared the structures of their most conserved sites. The MACAW program was used for discovering regions of statistically significant local similarity between sequences of different families. Analysis of 20 known sequences of family GH68 (Table I) revealed that they consist of 20 conserved regions (L1–L20) separated by spacers of variable length (Fig. 3, line 6). Eight of these regions are new, and the others were described earlier (Fig. 3, lines 2–5).[32–35] It should be noted that all common conserved regions of families GH32 and GH68[19] are highly conserved in levansucrases (Fig. 3, lines 6 and 7).

Comparison of 32 representative sequences of families GH43 and GH68 by the MACAW program revealed six common conserved regions (N1–N6) that are grouped into three clusters in the primary structure (Fig. 3, line 8). These regions are also conserved in the families GH32, GH62, and GHLP (Figs. 1 and 2). Analysis of PSI-BLAST alignment of N-terminal parts of the sequences revealed an additional homologous region that is conserved in families GH32, GH43, and GH62 (Fig. 4).

Several representative sequences of the each protein family were studied by hydrophobic cluster analysis. A closer examination of the conserved clusters of hydrophobic residues (V, I, L, F, M, W, and Y) in sequences of each family revealed three conserved hypothetical β-strands in the homologous regions for proteins of all five families. The three hydrophobic clusters locate in conserved regions N2, N4, and N5 (Figs. 1 and 2). These β-strands coincide with the ones proposed earlier by Pons et al.[36]

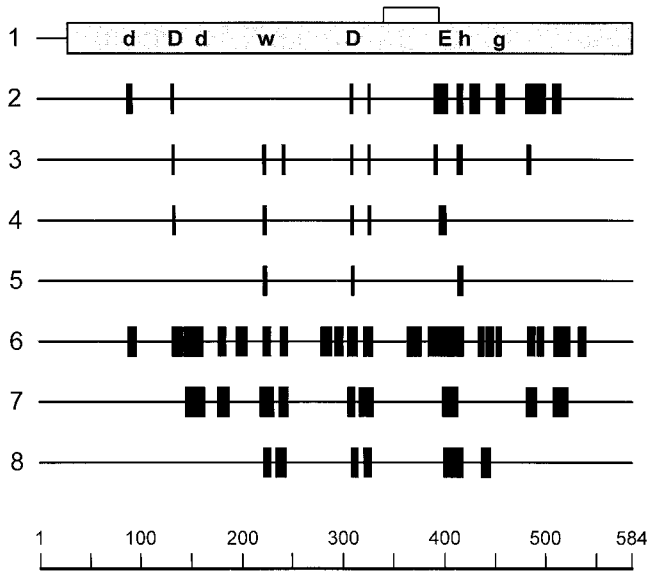

Fig. 3. Schematic positions of conserved regions in levansucrases. The scale corresponds to residue numbering in *G. diazotrophicus* levansucrase sequence. **1:** Positions of highly conserved amino acid residues discussed in the text. Capital letters correspond to proposed components of the active center. Mature form of the enzyme[50,51] is shown by box. Bracket indicates a disulfide bond. **2:** Regions (1–10) conserved in Proteobacteria levansucrases.[33] **3:** Regions (I–VIII) conserved in levansucrases.[33] **4:** Conserved motifs (I–V) containing acidic residues in levansucrases.[34] **5:** Regions (I–III) conserved in levansucrases.[35] **6:** Regions (L1–L20) conserved in 20 known levansucrases (present work). **7:** Conserved segments (D1–D9) of β-fructosidases from families GH32 and GH68.[19] **8:** Regions (N1–N6) conserved in glycosidases of families GH43 and GH68 (present work).

DISCUSSION

The families of protein sequences analyzed here possess several common conserved regions. One of them is located in the N-terminal part of the sequences (Fig. 4). This region includes the highly conserved Asp-Pro amino acid pair. The Asp residue was shown to be a nucleophile in the active center of *Saccharomyces cerevisiae* SUC2 invertase (SwissProt accession number P00724).[37] This amino acid pair and several surrounding residues are known as the "β-fructosidase motif"[38] and comprise a consensus pattern of family GH32 in PROSITE.[39] A frame shift in this site of *A. niger* invertase sequence (O13388) changed its activity for a fructosyltransferase.[27] The His residue (Fig. 4) is highly conserved in sequences of family GH32, and it was predicted that this residue locates in the active site.[36] However, side-directed mutagenesis showed that it is not involved in the catalytic process directly.[40] Another Asp residue in this region (Fig. 4) is highly conserved in sequences of families GH32 and GH43. This Asp and surrounding residues correspond to conserved "Asp box" of sialidases (family GH33, clan GH-E) and some other carbohydrate active enzymes, according to Rothe et al.[41] So, this long N-terminal region consists of three conserved

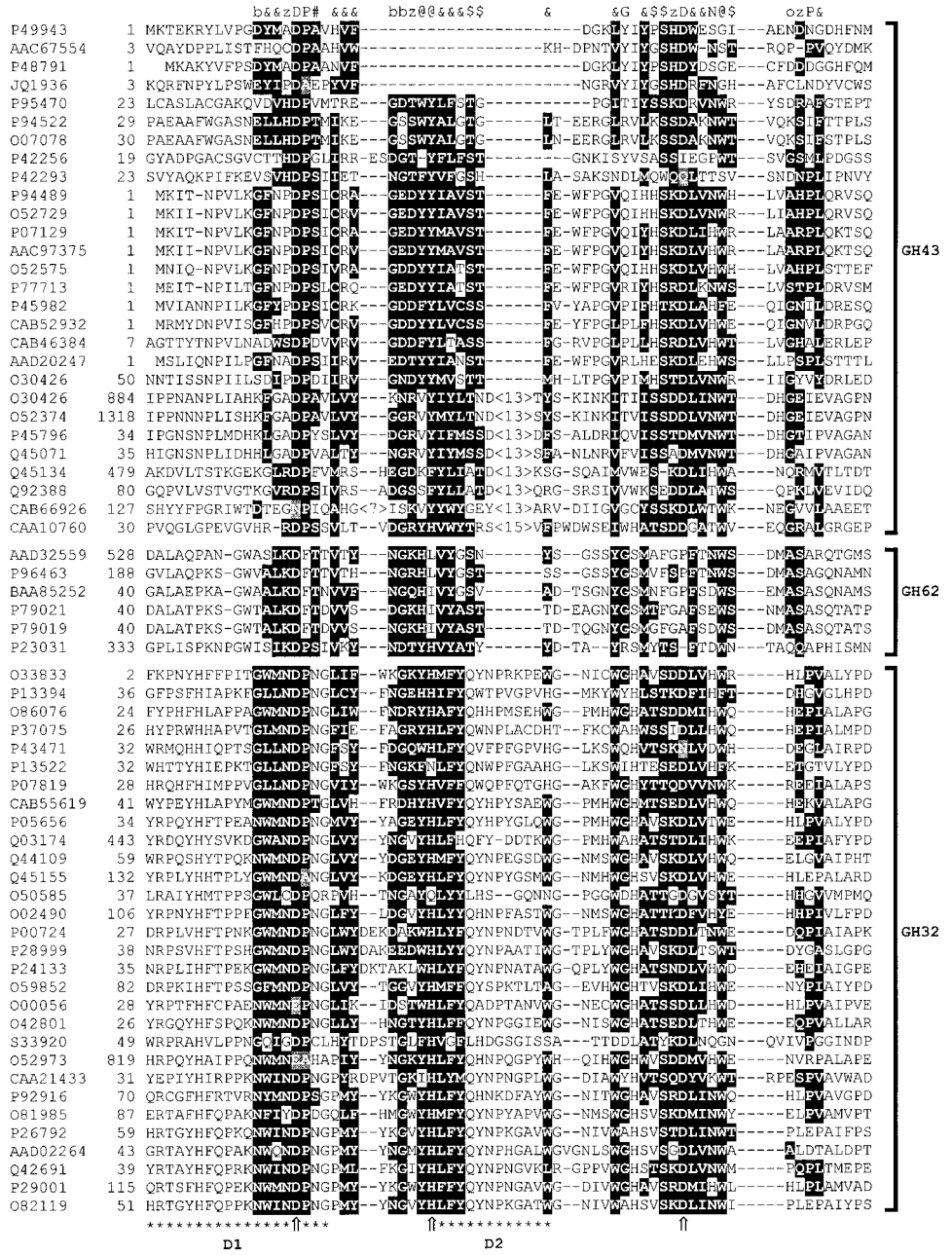

Fig. 4. Multiple sequence alignment of the sequences analyzed (fragment 1). Position of the first shown residue in each sequence and length of the variable spacer are indicated. At the top of the figure, residues conserved in several sequence families are indicated: & indicates a hydrophobic residue (I, L, V, M, F, Y, W, A, or C); @ indicates an aromatic residue (F, Y, W, or H); $ indicates a hydroxyl-containing residue (S or T); o indicates a small residue (G, S, or A); z indicates a polar residue (D, E, N, Q, H, K, R, S, or T); # indicates A, S, or T; b indicates D, E, N, Q, or G. The corresponding residues are highlighted in sequences (the trivial substitutions D/E/N/Q or P/A at the highly conserved sites are highlighted on a gray background). At the bottom of the figure, conserved regions D1 and D2 of β-fructosidases[19] are indicated by asterisks in the GH32 sequences. Conserved residues discussed in the text are indicated by arrows. The sequences are from the current sequence databases with the accession number for each sequence indicated in the leftmost column (for origin of the sequences see Table I). Family belongings of sequences are indicated in the most right.

sites (around two Asp and His residues of the yeast invertase) separated by the spacers of variable length (Fig. 4). This region can be easily identified in sequences of families GH32, GH43, and GH62. However, this region has a low degree of similarity with levansucrases. In their N-terminal part there are three conserved Asp residues; only one of them is invariant in all GH68 sequences (Fig. 5). Earlier we considered the third Asp as homologue of the invertase nucleophile (conserved segment D1 of families GH32 and GH68; Fig. 3, line 7; Figs. 4 and 5).[19] However, now it is clear that the invariant Asp (the second one) should be considered as its homologue. Hydrophobic cluster analysis of levansucrase sequences and several representatives of plant-type (subfamily 32d; Table I) and

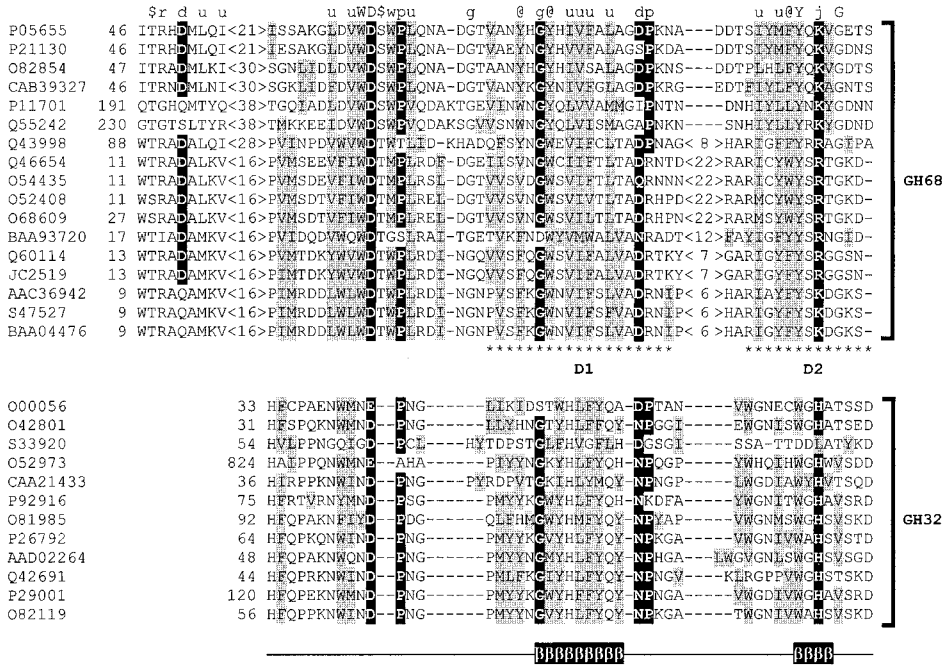

Fig. 5. Multiple sequence alignment of β-fructosidase sequences from families GH68 and GH32 (N-terminal fragment). At the top of the figure, residues conserved in sequences of family GH68 are indicated: u indicates a bulky hydrophobic residue (I, L, V, M, F, Y, or W); @ indicates an aromatic residue (F, Y, or W); $ indicates a hydroxyl-containing residue (S or T); j indicates a positively charged residue (K, R, or H). Invariant residues are shown by capital letters. Three Asp residues shown on Figure 3 (line 1) are highlighted. Residues conserved in both families are also highlighted (P, G, D/E/N/Q, or R/K/H). Conserved regions D1 and D2 of β-fructosidases[19] in the levansucrase sequences are indicated by asterisks. The bulky hydrophobic residues in the homologous parts of both families are shaded. At the bottom of the figure, two β-strands proposed by Pons et al.[36] for *Vigna radiata* invertase (P29001) are indicated. Other designations are the same as for Figure 4.

fungal β-fructosidase sequences (plant and *Aspergillus* invertases were shown to be the most similar group of family GH32 to levansucrases[36]) revealed three hydrophobic clusters in N-terminal region of sequences in both families (Fig. 5). These clusters probably correspond to β-strands in the secondary structure; two of these β-strands were proposed earlier for plant invertases by Pons et al.[36]

The other conserved regions are grouped into three clusters (Fig. 3, line 8). The first cluster corresponds to the "sucrose box," which was described as the homologous site of sucrases and levansucrases (families GH32 and GH68).[42] This cluster consists of two conserved regions (N1 and N2), separated by a short spacer of variable length (Fig. 1). It was found in four families: GH32, GH43, GH62, and GH68. The conserved region N1 has, at a conserved position, an aromatic residue (Phe, Trp, or Tyr), which is followed in enzymes with the β-fructosidase activity by a conserved hydroxyl-containing residue (Ser or Thr), invariant Gly residue, and a conserved hydroxyl-containing residue. The conserved region N2 contains a cluster of hydrophobic residues, which probably corresponds to a β-strand.

The second cluster also consists of two conserved regions (N3 and N4), separated by a spacer of variable length (Fig. 1). The region N3 includes a highly conserved Asp-Pro amino acid pair. In the case of sequences of families GH32 and GH68, this pair is preceded by the invariant Arg residue. This region is the most conserved site of these two families, and we proposed earlier that the Asp residue is a component of the β-fructosidase active center.[19] Recently, it was supported by site-directed mutagenesis of *Gluconacetobacter diazotrophicus* levansucrase: the Asp→Asn substitution affects sucrose hydrolysis, but not enzyme specificity.[34] This Asp was proposed as a possible catalytic residue in sequences of GH43 family.[43] The Asp residue is an invariant in all families of the superfamily except the GHLP one, where at this position an Asn residue is located. The region N4, like the region N2, contains a cluster of hydrophobic residues. This region also was proposed as a hypothetical β-strand.[36]

In all five families, the conserved region N5 includes an invariant Glu residue (with exception of AAD36300 and P71783) and a cluster of hydrophobic residues. In the case of families GH32 and GH62, these two structural elements are separated by a highly variable spacer (Fig. 2). On the basis of the site-directed mutagenesis it was proposed that this Glu residue is an acid/base catalyst in the active center of *S. cerevisiae* SUC2 invertase.[40] The sulfhydryl group of the conserved Cys residue preceded by the Glu in GH32 glycosidases is also essential for the enzymatic activity, its replacement leads to about fourfold reduction in K_{cat}.[40] The hydrophobic cluster consists, for the most part, of aromatic and, in particular, Tyr residues. The corresponding hypothetical β-strand is usually preceded by Gly residue, except sequences of family GH32. The region N5 of levansucrases includes a conserved positively charged residue (Arg or His; Fig. 2) essential for levanpolymerase activity.[44] The mutation leading to inactivation of *B. stearothermophilus* levansucrase (P94468) was also localized in the region N5.[30] This region in *G. diazotrophicus* levansucrase is preceded by Cys residue involved in a disulfide bridge (Fig. 3, line 1).

The conserved region N6 includes conserved Gly and an aromatic (Phe, Trp, or Tyr) residues. The region N6 is followed by a Gly residue invariant in family GH68 (Fig. 3, line 1), which is essential for secretion efficiency and folding process of *B. subtilis* levansucrase.[45] The regions N5 and N6 comprise the third conserved cluster, which we earlier described in sequences of families GH43 and GH68.[20]

TABLE II. Families of the Furanosidase (β-Fructosidase) Superfamily

Family[a]	GH32	GH43	GH62	GH68	GHLP
Clan[a]	GH-J	GH-F	GH-F	GH-J	None
Known enzymatic activities[b]	EC 2.4.1.99 EC 2.4.1.100 EC 2.4.1.x EC 3.2.1.7 EC 3.2.1.26 EC 3.2.1.65 EC 3.2.1.80	EC 3.2.1.8 EC 3.2.1.37 EC 3.2.1.55 EC 3.2.1.99	EC 3.2.1.55	EC 2.4.1.10 EC 3.2.1.26	Not known
Molecular mechanism	Retaining	Inverting	Not known	Retaining	Not known
Origin	Eukaryota Euglenozoa Fungi Parabasalidea Viridiplantae Eubacteria Cytophagales Firmicutes Fusobacteria[f] Proteobacteria Thermotogales	Eukaryota Fungi Metazoa[h] Viridiplantae[i] Eubacteria Cytophagales Firmicutes Proteobacteria	Eukaryota Fungi Eubacteria Firmicutes Proteobacteria	Eubacteria Firmicutes Proteobacteria	Eubacteria Aquificales Firmicutes Thermotogales Archaea Crenarchaeota Euryarchaeota
N-terminal conserved region[c]	NDPNG	DP	LKDF	WD$WP	None
Conserved region N1[d]	@G[g]	WAP	WVY	EWSG$	None
Conserved region N2[d]	Variable	Hydrophobic	Variable	Hydrophobic	None
Conserved region N3[d]	FRDP	DP	IDQ	FRDP	@NP
Conserved region N5[e]	@ECP	ExP	LFEA	ERP	EDPR
Spacer between Glu and hydrophobic cluster in conserved region N5[e]	Variable	Conserved[j]	Divergent conserved	Conserved	Conserved
Conserved region N6[e]	Variable Gx@	Conserved[k] GP@	Divergent conserved GxWT	Conserved GPY	Variable GxY

[a]According to the classification of glycosyl hydrolases.[10,11] GHLP family is a glycosyl hydrolase like protein family described in the text.
[b]See note "g" in Table I.
[c]See Figs. 4 and 5.
[d]See Fig. 1.
[e]See Fig. 2.
[f]N-terminal sequence of *Fusobacterium mortiferum* sucrase has been published.[54]
[g]@ indicates an aromatic residue (F, Y, or W) and $ indicates a hydroxyl-containing residue (S or T).
[h]Partly sequenced *Homo sapiens* ORF (GenPept accession number ACC34952) is homologous to the C-terminal domain of sequences from subfamily 43c (it is not considered in the text).
[i]Only an ORF from *Arabidopsis thaliana* (CAB66926 and CAA10760, subfamily 43f) is known.
[j]Variable in sequences of subfamily 43e.
[k]Absent in sequences of subfamily 43e.

Statistically significant sequence similarity of proteins of the five families allows us to propose a common protein folding and structure of the active center. However, the final decision about degree of similarity of the proteins of these families should be made after receiving experimental data on their 3D structure. Most probably, all proteins of the superfamily have a trio of conserved carboxylic acids in the active site (such kind of active site structure has been shown for some glycosidases[12,46]). They are Asp and Glu residues corresponding to the yeast invertase nucleophile and acid/base catalyst (Figs. 2 and 4) and the Asp residue of conserved region N3 (Fig. 1). Two Cys residues located near the second Asp and the Glu residues in the *G. diazotrophicus* levansucrase sequence (Fig. 3, line 1) comprise a disulfide bridge. There is a 25-amino acid deletion in the spacer between the conserved regions N4 and N5 of *Streptococcus mutans* invertase (P13522). These two facts indicate that the second Asp residue (region N3) can be located at the active site together with the first Asp and Glu. It should be noted that these three carboxylic residues are usually followed by Pro. Both Asp-Pro amino acid pairs are conserved in families GH32, GH43, and GH68; however, in family GH68 the first of them is modified as Asp-Xaa-Xaa-Pro (Fig. 5). In both positions, Pro residues are substituted by Ala in some sequences (Figs. 1 and 4). Also, there is some similarity between residues surrounding Asp-Pro pairs. It suggests similar chemical properties for the two Asp residues.

The data presented in this article allow us to include into the β-fructosidase superfamily, in addition to the families of clan GH-J (GH32 and GH68), families of clan GH-F (GH43 and GH62), which have the α-L-arabinofuranosidase activity present but show no β-D-fructofuranosidase activity (Table II). It looks reasonable to

rename this superfamily as the furanosidase superfamily. It is shown for the first time that a glycosyl hydrolase superfamily can include enzymes with both inversion and retention mechanism of action. (A superfamily that consists of enzymes with both mechanisms of action was proposed earlier.[47,48] However, the proteins of families GH19, GH22, GH23, GH24, and GH46, which were included into the latter superfamily, had similarities at the levels of secondary and tertiary but not primary structures.)

We suggest the term "superfamily" for designating the group of proteins with a higher hierarchical level than clan. In contrast, clan includes only families of structurally related proteins having the same mechanism of action. So, we still consider GH-F and GH-J as two separate clans of the furanosidase superfamily. Despite the fact that all known glycoside hydrolases with a high degree of homology have the same mechanism of hydrolysis and that enzymes with different mechanisms have a different degree of separation of two key carboxyl groups in the active center (about 5 and 10 Å between a nucleophile and an acid/base catalyst for retaining and inverting enzymes, respectively), it was shown that single amino acid substitution can change the mechanism.[12,49] The GHLP family also belongs to the furanosidase superfamily on the basis of sequence similarity. This allows us to propose that hypothetical proteins of this family can have a glycosidase activity (EC 3.2.1.x) and, most probably, they act on a furanoside residue (fructose, arabinose, ribose, etc.). Genes of proteins of the superfamily have been found in the genomes of all main groups of organisms except viruses (Table II), which suggests that they have a very early evolution origin.

REFERENCES

1. Enzyme Nomenclature. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes. San Diego: Academic Press; 1992. 862 p.
2. Henrissat B. A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem J 1991;280:309–316.
3. Coutinho PM, Henrissat B. The modular structure of cellulases and other carbohydrate-active enzymes: an integrated database approach. In: Ohmiya K, Hayashi K, Sakka K, Kobayashi Y, Karita S, Kimura T, editors. Genetics, biochemistry and ecology of cellulose degradation. Tokyo: Uni Publishers Co.; 1999. p 15–23.
4. Coutinho PM, Henrissat B. Carbohydrate-active enzymes: an integrated database approach. In: Gilbert HJ, Davies G, Henrissat B, Svensson B, editors. Recent advances in carbohydrate bioengineering. Cambridge, UK: The Royal Society of Chemistry; 1999. p 3–12.
5. Davies G, Henrissat B. Structures and mechanisms of glycosyl hydrolases. Structure 1995;3:853–859.
6. Henrissat B, Davies G. Structural and sequence-based classification of glycoside hydrolases. Curr Opin Struct Biol 1997;7:637–644.
7. Henrissat B, Bairoch A. New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem J 1993;293:781–788.
8. Himmel ME, Karplus PA, Sakon J, Adney WS, Baker JO, Thomas SR. Polysaccharide hydrolase folds diversity of structure and convergence of function. Appl Biochem Biotech 1997;63–65:315–325.
9. Henrissat B, Bairoch A. Updating the sequence-based classification of glycosyl hydrolases. Biochem J 1996;316:695–696.
10. Bairoch A. SWISS-PROT Protein Sequence Data Bank. 2000 (http://www.expasy.ch/cgi-bin/lists?glycosid.txt).
11. Coutinho PM, Henrissat B. Carbohydrate-Active Enzymes server. 2000 (http://afmb.cnrs-mrs.fr/~pedro/CAZY/db.html).
12. McCarter JD, Withers SG. Mechanisms of enzymatic glycoside hydrolysis. Curr Opin Struct Biol 1994;4:885–892.
13. Tomme P, Warren RAJ, Gilkes NR. Cellulose hydrolysis by bacteria and fungi. Adv Microb Physiol 1995;37:1–81.
14. Kuriki T, Imanaka T. The concept of the α-amylase family: structural similarity and common catalytic mechanism. J Biosci Bioeng 1999;87:557–565.
15. Gebler J, Gilkes NR, Claeyssens M, et al. Stereoselective hydrolysis catalyzed by related β-1,4-glucanases and β-1,4-xylanases. J Biol Chem 1992;267:12559–12561.
16. Henrissat B, Romeu A. Families, superfamilies and subfamilies of glycosyl hydrolases. Biochem J 1995;311:350–351.
17. Henrissat B. Glycosidase families. Biochem Soc Trans 1998;26:153–156.
18. Mian IS. Sequence, structural, functional, and phylogenetic analyses of three glycosidase families. Blood Cells Mol Dis 1998;24:83–100.
19. Naumov DG, Doroshenko VG. β-Fructosidases: a new superfamily of glycosyl hydrolases. Mol Biol (Engl Transl) 1998;32:761–766.
20. Naumoff DG. Conserved sequence motifs in levansucrases and bifunctional β-xylosidases and α-L-arabinases. FEBS Lett 1999; 448:177–179.
21. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
22. Schuler GD, Altschul SF, Lipman DJ. A workbench for multiple alignment construction and analysis. Proteins Struct Funct Genet 1991;9:180–190.
23. Callebaut I, Labesse G, Durand P, et al. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. Cell Mol Life Sci 1997;53:621–645.
24. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
25. Mercier RW, Chaivisuthangkurar P, Gogarten JP. Invertase encoding cDNA from oat [letter]. Plant Mol Biol 1993;23:229–230.
26. Bezzate S, Steinmetz M, Aymerich S. Cloning, sequencing, and disruption of a levanase gene of *Bacillus polymyxa* CF43. J Bacteriol 1994;176:2177–2183.
27. Somiari RI, Brzeski H, Tate R, Bieleck S, Polak J. Cloning and sequencing of an *Aspergillus niger* gene coding for β-fructofuranosidase. Biotech Lett 1997;19:1243–1247.
28. Liebl W, Brem D, Gotschlich A. Analysis of the gene for β-fructosidase (invertase, inulinase) of the hyperthermophilic bacterium *Thermotoga maritima*, and characterisation of the enzyme expressed in *Escherichia coli*. Appl Microbiol Biotechnol 1998;50:55–64.
29. Naumoff DG. Levanase gene sequence from strain *Bacillus* sp. L7 [letter]. FEMS Microbiol Lett 1998;164:227–228.
30. Naumoff DG. Homologous locus of *Bacillus subtilis* and *Bacillus stearothermophilus* genomes containing levansucrase and levanase genes. Mol Biol (Engl Transl) 1999;33:173–176.
31. Naumoff DG, Doroshenko VG, Livshits VA. Sequencing of a fragment of the sucrase gene from *Lactobacillus plantarum*. Mol Biol 1998;32:373 (in Russian).
32. Song K-B, Joo H-K, Rhee S-K. Nucleotide sequence of levansucrase gene (*levU*) of *Zymomonas mobilis* ZM1 (ATCC10988). Biochim Biophys Acta 1993;1173:320–324.
33. Arrieta J, Hernández L, Coego A, et al. Molecular characterization of the levansucrase gene from the endophytic sugarcane bacterium *Acetobacter diazotrophicus* SRT4. Microbiology 1996;142: 1077–1085.
34. Batista FR, Hernández L, Fernández JR, et al. Substitution of Asp-309 by Asn in the Arg-Asp-Pro (RDP) motif of *Acetobacter diazotrophicus* levansucrase affects sucrose hydrolysis, but not enzyme specificity. Biochem J 1999;337:503–506.
35. Kurimoto M, Tsusaki K, Kubota M, Fukuda S, Tsujisaka Y. Cloning and sequencing of the β-fructofuranosidase gene from *Bacillus* sp. V230. Biosci Biotech Biochem 1999;63:1107–1111.
36. Pons T, Olmea O, Chinea G, et al. Structural model for family 32 of glycosyl-hydrolase enzymes. Proteins Struct Funct Genet 1998;33: 383–395.
37. Reddy VA, Maley F. Identification of an active-site residue in yeast invertase by affinity labeling and site-directed mutagenesis. J Biol Chem 1990;265:10817–10820.

38. Sturm A, Chrispeels MJ. cDNA cloning of carrot extracellular β-fructosidase and its expression in response to wounding and bacterial infection. Plant Cell 1990;2:1107–1119.
39. Bairoch A. The PROSITE dictionary of sites and patterns in proteins, its current status. Nucleic Acids Res 1993;21:3097–3103.
40. Reddy A, Maley F. Studies on identifying the catalytic role of Glu-204 in the active site of yeast invertase. J Biol Chem 1996;271:13953–13958.
41. Rothe B, Rothe B, Roggentin P, Schauer R. The sialidase gene from *Clostridium septicum*: cloning, sequencing, expression in *Escherichia coli* and identification of conserved sequences in sialidases and other proteins. Mol Gen Genet 1991;226:190–197.
42. Sato Y, Kuramitsu HK. Sequence analysis of the *Streptococcus mutans scrB* gene. Infect Immun 1988;56:1956–1960.
43. Matsuo N, Kaneko S, Kuno A, Kobayashi H, Kusakabe I. Purification, characterization and gene cloning of two α-L-arabinofuranosidases from *Streptomyces chartreusis* GS901. Biochem J 2000;346:9–15.
44. Chambert R, Petit-Glatron M-F. Polymerase and hydrolase activities of *Bacillus subtilis* levansucrase can be separately modulated by site-directed mutagenesis. Biochem J 1991;279:35–41.
45. Petit-Glatron MF, Monteil I, Benyahia F, Chambert R. *Bacillus subtilis* levansucrase: amino acid substitutions at one site affect secretion efficiency and refolding kinetics mediated by metals. Mol Microbiol 1990;4:2063–2070.
46. White A, Rose DR. Mechanism of catalysis by retaining β-glycosyl hydrolases. Curr Opin Struct Biol 1997;7:645–651.
47. Holm L, Sander C. Structural similarity of plant chitinase and lysozymes from animals and phage: an evolutionary connection. FEBS Lett 1994;340:129–132.
48. Monzingo AF, Marcotte EM, Hart PJ, Robertus JD. Chitinases, chitosanases, and lysozymes can be divided into procaryotic and eucaryotic families sharing a conserved core. Nat Struct Biol 1996;3:133–140.
49. Kuroki R, Weaver LH, Matthews BW. Structure-based design of a lysozyme with altered catalytic activity. Nat Struct Biol 1995;2:1007–1011.
50. Betancourt L, Takao T, Hernandez L, Padron G, Shimonishi Y. Structural characterization of *Acetobacter diazotrophicus* levansucrase by matrix-assisted laser desorption/ionization mass spectrometry: identification of an N-terminal blocking group and a free-thiol cysteine residue. J Mass Spectrom 1999;34:169–174.
51. Hernández L, Arrieta J, Betancourt L, et al. Levansucrase from *Acetobacter diazotrophicus* SRT4 is secreted via periplasm by a signal-peptide-dependent pathway. Curr Microbiol 1999;39:146–152.
52. Gasparic A, Martin J, Daniel AS, Flint HJ. A xylan hydrolase gene cluster in *Prevotella ruminicola* B_14: sequence relationships, synergistic interactions, and oxygen sensitivity of a novel enzyme with exoxylanase and β-(1,4)-xylosidase activities. Appl Environ Microbiol 1995;61:2958–2964.
53. Wegener S, Ransom RF, Walton JD. A unique eukaryotic β-xylosidase gene from the phytopathogenic fungus *Cochliobolus carbonum*. Microbiology 1999;145:1089–1095.
54. Thompson J, Nguyen NY, Robrish SA. Sucrose fermentation by *Fusobacterium mortiferum* ATCC 25557: transport, catabolism, and products. J Bacteriol 1992;174:3227–3235.