ISSN 0026-8933, Molecular Biology, 2007, Vol. 41, No. 6, pp. 962–973. © Pleiades Publishing, Inc., 2007.
Original Russian Text © D.G. Naumoff, 2007, published in Molekulyarnaya Biologiya, 2007, Vol. 41, No. 6, pp. 1056–1068.

MATHEMATICAL AND SYSTEM BIOLOGY

UDC 575.852:577.124.2:577.152.321

Structure and Evolution of the Mammalian Maltase–Glucoamylase and Sucrase–Isomaltase Genes

D. G. Naumoff

Institute of Genetics and Selection of Industrial Microorganisms, Moscow, 117545 Russia;
e-mail: daniil_naumoff@yahoo.com

Received October 27, 2006

Accepted for publication April 6, 2007

Abstract—Maltase–glucoamylase and sucrase–isomaltase are two glycosydases responsible for starch digestion in human. Their evolutionary history was studied by comparing the amino acid sequences of these enzymes from several mammals and their orthologs from other chordates. The two glycosydases are paralogs and contain catalytic domains of the GH31 family. A common evolutionary precursor of their genes arose via a tandem duplication. As a consequence, sucrase–isomaltase consists of two homologous parts. The maltase–glucoamylase gene experienced several additional duplications, whose number varies among mammals. Its locus harbors four to seven tandem repeats, each coding for an amino acid sequence similar to the two parts of sucrase–isomaltase.

DOI: 10.1134/S0026893307060131

Key words: glycoside hydrolase, starch utilization, GH31 family, α-glucosidase, paralog, protein family, protein phylogenetic tree, domain structure, gene duplication, enzyme classification, multiple sequence alignment, gene annotation

INTRODUCTION

Maltase–glucoamylase [EC 3.2.1.20 and 3.2.1.3] and sucrase–isomaltase [EC 3.2.1.48 and 3.2.1.10] are two glycoside hydrolases (glycosidases) with complementary activities [1–3]. Sucrase–isomaltase digests branched starch linkages in the mammalian intestine. Maltase–glucoamylase digests linear regions to yield glucose at the last step of starch hydrolysis. The two enzymes are paralogs and each contain two homologous catalytic domains belonging to glycoside hydrolase family 31 (GH31) [4].

Genes for GH31-family proteins are found in the overwhelming majority of organisms, including bacteria, archaea, and eukaryotes. Several paralogs are encoded by one genome in many cases. For instance, eight GH31 proteins are known in human [5]. Other mammals have a similar repertoire of GH31 proteins. Most genes of GH31 proteins are open reading frames that have not been experimentally characterized. Biochemically characterized GH31 proteins possess various enzymatic activities [5] and include glucosyltransferases [EC 2.4.1.–], isomaltosyltransferases [EC 2.4.1.–], glucoamylases [EC 3.2.1.3], isomaltases [EC 3.2.1.10], α-glucosidases [EC 3.2.1.20], sucrases [EC 3.2.1.48], α-1,3-glucosidases [EC 3.2.1.84], α-xylosidases [EC 3.2.1.–], and α-glucan lyases [EC 4.2.2.13]. The 3D structure has recently been solved for the first two GH31 proteins, *Escherichia coli* YicI α-xylosidase [6] and *Sulfolobus solfataricus* MalA α-glucosidase [7] (PDB IDs 1WE5 and 2G3M). Each of these proteins consists of four main domains: N, A, C, and D (Fig. 1a). Catalytic domain A belongs to the GH31 family, has a $(\beta/\alpha)_8$-barrel (TIM barrel-type) structure, and occupies the central region of the total amino acid sequence. One (B) or two (B and B') additional subdomains are inserted in domain A in YicI and MalA, respectively. Domains N, C, and D have a β-sandwich structure.

Sequence comparisons have revealed that glycosidases of the GH27, GH31, and GH36 families have related catalytic domains [7–14]; accordingly, the three families have been assigned to one α-galactosidase superfamily [9, 11, 13, 14]. More distant evolutionary relationships have been observed with the GH13 and GH97 families [6, 7, 10, 13–19]. Proteins of these families have a $(\beta/\alpha)_8$-barrel catalytic domain and cleave the axial glycopyranoside bond, retaining the optical configuration of the hydrolysis product [5]. A common evolutionary origin is possible to trace for almost all glycosidase families with the TIM barrel-type catalytic domain [14].

The objective of this work was to study the structure and evolution of the maltase–glucoamylase and sucrase–isomaltase genes of human and other mam-

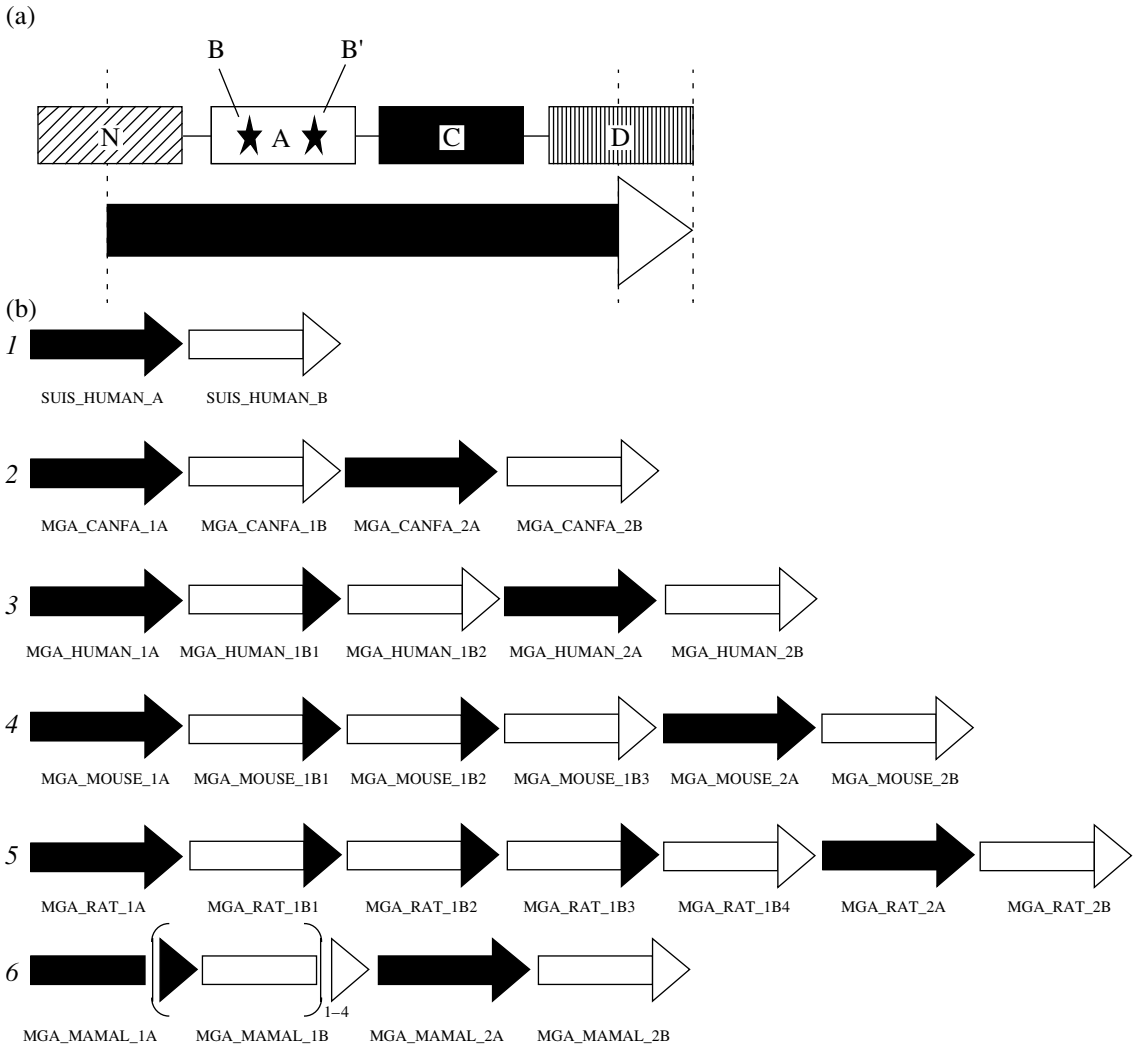


Fig. 1. (a) Domain structure of GH31 proteins and (b) module structure of some proteins examined in this work. N, A, C, and D are the main domains of *Sulfolobus solfataricus* α-glucosidase MalA (PDB ID 2G3M). Catalytic domain A has the $(\beta/\alpha)_8$-barrel structure and contains additional subdomains B and B'. The GH31 module is shown with an arrow. Its two parts are filled and open; their correspondence to the MalA domains is shown with dashed lines: the C-terminal region of the module contains the second half of the last domain (D). Enzymes: *1*, human sucrase–isomaltase (structurally similar to other mammalian sucrases–isomaltases); *2*, dog maltase–glucoamylase (similar to bovine maltase–glucoamylase); *3*, human maltase–glucoamylase (similar to chimpanzee and macaque maltases–glucoamylases); *4*, mouse maltase–glucoamylase; *5*, rat maltase–glucoamylase; *6*, integral scheme of mammalian maltase–glucoamylase (the fragment subject to multiple tandem duplications is shown in parentheses). Type A modules are filled; type B modules are open. In hybrid type B modules, the C end is filled. The modules are designated as in Table 2.

mals by comparing the genome sequencing data. Preliminary results of this work have been reported in the proceedings of the XXIII International Carbohydrate Symposium [20].

DATA ANALYSIS

The amino acid and nucleotide sequences were extracted from the NCBI database (http://www.ncbi.nlm.nih.gov/) (Tables 1, 2). The list of GH31 proteins available from the CAZy site (http://www.cazy.org) was used as a basis. The list was supplemented by screening the NCBI database with the blastp software program (http://www.ncbi.nlm.nih.gov/blast). Homologs of the sucrase–isomaltase genes were sought in the macaque and ascidian genomes with the help of the tblastn software program (http://www.ncbi.nlm.nih.gov/blast). Multiple sequence alignments and subsequent phylogenetic analysis were performed with the amino acid sequences deduced from the nucleotide sequences. The putative exon–intron structure was inferred from the NCBI data. However, when multiple sequence alignment revealed extended gaps in particular amino acid sequences or a local lack of similarity to other sequences, the corresponding nucleotide sequences were analyzed in the three reading frames and the

Table 1. Human proteins possessing GH31 domains (according to the CAZy database)

GenPept acc. no.	Protein	Gene	Chromosome	Protein size, residues	Identity to CAA68763.1, %	Identity to AAH59406.1, %	Subfamily
CAA45140.1	Intestinal sucrase–isomaltase	*SI*	3	1827			
	N-terminal part (sucrase)			936	41	27	31a
	C-terminal part (isomaltase)			891	38	25	31a
AAC39568.2	Intestinal maltase–glucoamylase	*MGAM*, *MGA*	7	1857			
	N-terminal part (maltase)			959	43	27	31a
	C-terminal part (glucoamylase)			898	38	25	31a
AAI11974.1	Unknown*	*LOC93432*	7	482	41	22	31a
BAD18495.1	Unknown*	–	(7)	646	43	28	31a
CAA68763.1	Acid (lysosomal) α-glucosidase	*GAA*	17	952	100	29	31a
AAH59406.1	Neutral α-glucosidase C	*GANC*, *FLJ00088*	15	914	30	100	31b
AAF66685.1	Neutral α-glucosidase AB	*KIAA0088*, *GANAB*	11	966	32	49	31b
AAH70098.1	Unknown*	*LOC57462*, *KIAA1161*	9	714	23	22	31c

Note: The name of protein BAD18495.1 and the chromosomal location of its gene are missing in the GenPept database. The sizes of the two individual parts are given for maltase–glucoamylase and sucrase--isomaltase. The identity of CAA68763.1 and AAH59406.1 with other sequences was estimated from the pairwise sequence alignment, which was yielded by PSI-BLAST (http://www.ncbi.nlm.nih.gov/blast/) after two rounds with the two proteins used as queries.

* Proteins are marked as unknown when the corresponding sequence is known only at the mRNA level.

exon–intron structure was corrected to improve the similarity of the amino acid sequences in question. This was performed using the tblastn program, which finds the best pairwise sequence alignment of a query protein with nucleotide sequences read in the three frames.

Multiple amino acid sequence alignments were manually constructed using the BioEdit program (http://www.mbio.ncsu.edu/BioEdit/bioedit.html), using the pairwise sequence alignments constructed with the PSI-BLAST and tblastn programs on the NCBI server. Multiple sequence alignment was performed with protein fragments homologous to the C-terminal region (starting from residue 937) of human sucrase–isomaltase (GenPept, CAA45140.1). After removing the most variable regions, the alignment was used to construct phylogenetic trees with the help of the PROTPARS (protein sequence parsimony method, MP) and NEIGHBOR (neighbor-joining method, NJ) programs from the PHYLIP software package (http://evolution.gs.washington.edu/phylip.html). The statistical reliability of nodes was evaluated by bootstrap analysis with 1000 pseudoreplicates for each tree. Trees were displayed using the TreeView Win32 program (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html). Only NJ trees are shown. Glycosidases were classified into families as on the CAZy site; organisms were classified as on the NCBI Taxonomy Homepage (http://www.ncbi.nlm.nih.gov/Taxonomy/).

RESULTS

Eight human proteins (Table 1) are now assigned to the GH31 family in the CAZy database [5]. Of these, maltase–glucoamylase and sucrase–isomaltase each have two GH31 domains and the other proteins each harbor one GH31 domain. Pairwise sequence comparisons of these proteins showed that their similarity is not restricted to the catalytic domain, having the $(\beta/\alpha)_8$-barrel structure and belonging to the GH31 family, but is extended almost over the total protein length. The common homologous region (corresponding to region 937–1827 of human sucrase–isomaltase; GenPept, CAA45140.1) covers, at least partly, all four domains of MalA and YicI [6, 7]. This region of the human proteins and homologous regions of GH31 proteins from other organisms are hereafter referred to as the GH31 module (Fig. 1a). The module included the entire Glyco_hydro_31 domain from the Pfam database (http://www.sanger.ac.uk/Software/Pfam/) and about 150 additional residues. In some cases, the GH31 module was incomplete because of the lack of the N- or C-terminal region or deletions from the internal region. Human maltase–glucoamylase and sucrase–isomaltase each consist of two tandem GH31

Table 2. GH31 modules subject to phylogenetic analysis

Module	Organism	GenBank acc. no.	Module size, residues	Comments
SUIS_HUMAN_A	*Homo sapiens*	X63597.1	866	
SUIS_HUMAN_B	"	X63597.1	891	
MGA_HUMAN_1A	"	NT_007914.14	865	
MGA_HUMAN_1B1*	"	NT_007914.14	896	A/B
MGA_HUMAN_1B2*	"	NT_007914.14	898	
MGA_HUMAN_2A*	"	NT_007914.14	862	
MGA_HUMAN_2B*	"	NT_007914.14	903	
SUIS_PANTR_A*	*Pan troglodytes*	XM_526371.2	822***	
SUIS_PANTR_B*	"	XM_526371.2	854***	
MGA_PANTR_1A*	"	NW_001238096.1	863	
MGA_PANTR_1B1*	"	NW_001238096.1	753***	A/B
MGA_PANTR_1B2*	"	NW_001238096.1	771***	
MGA_PANTR_2A	"	NW_001238096.1	862	
MGA_PANTR_2B	"	NW_001238096.1	903	
SUIS_ MACMU_A**	*Macaca mulatta*	AANU01251577, AANU01251578, AANU01251579, AANU01251580, AANU01146165, AANU01251581	838***	
SUIS_ MACMU_B**	"	AANU01251581, AANU01251582, AANU01251583, AANU01251584	890	
MGA_MACMU_1A	"	XM_001083672.1	865	
MGA_MACMU_1B1	"	XM_001083672.1	910	A/B
MGA_MACMU_1B2*	"	XM_001083672.1, XM_001118718.1, XM_001083773.1	898	
MGA_MACMU_2A*	"	XM_001083890.1, AANU01114149.1, XM_001083998.1	862	
MGA_MACMU_2B*	"	XM_001083998.1, AANU01114152.1	907	
SUIS_CANFA_A	*Canis familiaris*	XM_545265.2	866	
SUIS_CANFA_B	"	XM_545265.2	889	
MGA_CANFA_1A	"	NW_876260.1	865	
MGA_CANFA_1B	"	NW_876260.1	899	
MGA_CANFA_2A*	"	NW_876260.1	861	
MGA_CANFA_2B	"	NW_876260.1	903	
SUIS_BOSTA_A*	*Bos taurus*	XM_580476.2	825***	
SUIS_BOSTA_B*	"	XM_580476.2	886	
MGA_BOSTA_1A	"	NW_001015003.1	864	
MGA_BOSTA_1B*	"	NW_001015003.1, NW_931372.1	861***	
MGA_BOSTA_2A*	"	NW_931379.1	832***	
MGA_BOSTA_2B*	"	NW_931379.1	904	

Table 2. (Contd.)

Module	Organism	GenBank acc. no.	Module size, residues	Comments
SUIS_MOUSE_A*	*Mus musculus*	XM_143332.7	866	
SUIS_MOUSE_B*	"	XM_143332.7	891	
MGA_MOUSE_1A*	"	NT_039341.6	864	
MGA_MOUSE_1B1*	"	NT_039341.6	895	A/B
MGA_MOUSE_1B2*	"	NT_039341.6	894	A/B
MGA_MOUSE_1B3	"	NT_039341.6	894	
MGA_MOUSE_2A	"	NT_039341.6	865	
MGA_MOUSE_2B*	"	NT_039341.6	909	
SUIS_RAT_A	*Rattus norvegicus*	L25926.1	866	
SUIS_RAT_B*	"	L25926.1	890	
MGA_RAT_1A*	"	NW_047690.1	863	
MGA_RAT_1B1*	"	NW_047690.1	896	A/B
MGA_RAT_1B2	"	NW_047690.1	896	A/B
MGA_RAT_1B3*	"	NW_047690.1	884	A/B
MGA_RAT_1B4*	"	NW_047690.1	893	
MGA_RAT_2A	"	NW_047690.1	860	
MGA_RAT_2B*	"	NW_047690.1	852***	?
SUIS_SUNMU_A	*Suncus murinus*	AB011401.1	866	
SUIS_SUNMU_B	"	AB011401.1	892	
SUIS_RABIT_A	*Oryctolagus cuniculus*	M14046.1	866	
SUIS_RABIT_B	"	M14046.1	891	
ORF1_CHICK_A*	*Gallus gallus*	XM_422811.1	865	
ORF1_CHICK_B	"	XM_422811.1	893	A/B
ORF1_TETNI_A*	*Tetraodon nigroviridis*	CAAE01014985.1	819***	
ORF1_TETNI_B	"	CAAE01014985.1	760***	
ORF1_CIOIN_A**	*Ciona intestinalis*	AABS01000112.1	819***	–
ORF1_CIOIN_B**	"	AABS01000112.1	789***	–
ORF1_CIOSA_A**	*Ciona savignyi*	AACT01021410.1, AACT01018394.1	813***	–
ORF1_CIOSA_B**	"	AACT01021410.1	307***	–
ORF1_STRPU_A	*Strongylocentrotus purpuratus*	XM_792178.2	891	A
ORF1_STRPU_B	"	XM_792178.2	755	–

Notes: The modules are designated according to the classification proposed (see text). Two or more accession numbers are given when one sequence coding for the full-size module is unavailable from GenBank. Comments: A/B, hybrid modules, with most of the sequence corresponding to type B and the C-terminal part of the last domain corresponding to type A (Fig. 1); (?), the type B C-terminal part is fragmentary; (–), the same part is completely absent.

* The exon–intron boundaries of the corresponding gene fragments were changed as compared to those indicated in GenBank.

** The module gene is not annotated in GenBank (the amino acid sequence was deduced from the corresponding DNA sequence).

*** The size is indicated for a fragment rather than for the total protein.

modules. Pairwise sequence comparisons of all ten human GH31 modules showed that the modules substantially differed in the extent of their sequence similarity. The modules could be divided into three groups (a, b, and c), each including those with more than 30% sequence identity (Table 1). The three groups of the human proteins belong to the three major clusters of eukaryotic proteins on a phylogenetic tree of the GH31 family (unpublished data). It should be noted that a 30% sequence identity of amino acid sequences is regarded as one of the major criteria in dividing the glycosidase families into subfamilies

[12–14, 18, 21–24]. Thus, groups a, b, and c can be considered representing three subfamilies of GH31 modules (31a–31c, Table 1).

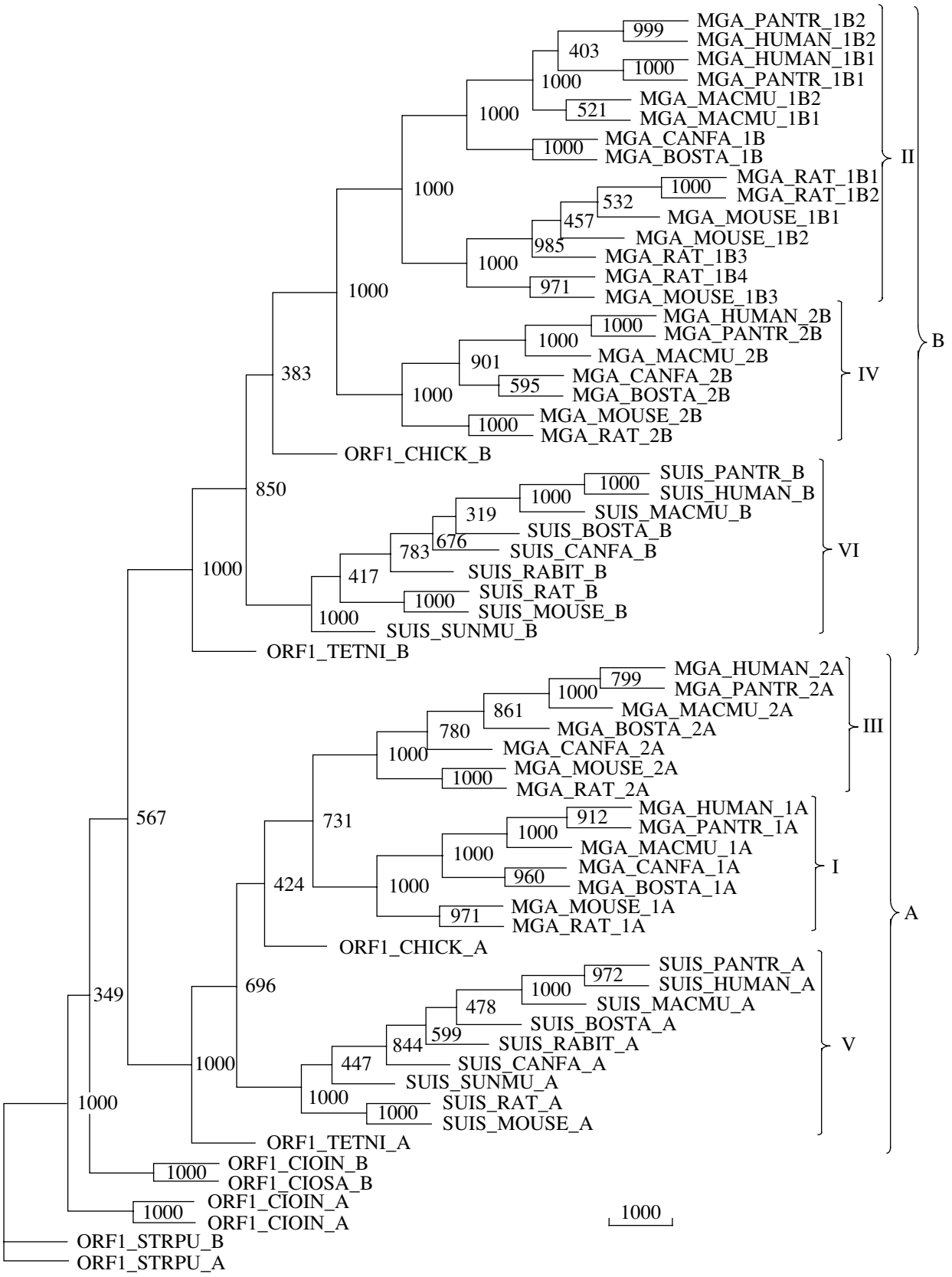
The human sucrase–isomaltase gene is on chromosome 3. The maltase–glucoamylase gene is on chromosome 7, along with two other genes (GenPept, AAI11974.1 and BAD18495.1) for proteins possessing group a GH31 modules (Table 1). Screening of the amino acid sequence database revealed many mammalian proteins with more than one GH31 module. For instance, rat *Rattus norvegicus* maltase–glucoamylase (GenPeptXP_231714.1) has seven modules. Its chimpanzee *Pan troglodytes* ortholog (GenPeptXP_519433.1) harbors three modules; the central one has rearrangements of several internal fragments, which can be explained by an evolutionarily recent mutation or a misassembly of the corresponding chromosome contig. Human proteins AAI11974.1 and BAD18495.1 correspond to one chimpanzee protein, GenPeptXP_519434.1, consisting of two GH31 modules. The mouse *Mus musculus* genome codes not only for sucrase–isomaltase (GenPeptXP_143332.6) but also for three other proteins (GenPeptXP_485746.1, XP_133071.2, and XP_487916.1), which do not result from alternative splicing and each harbor two GH31 modules. The new version of the mouse maltase–glucoamylase gene (GenPeptXP_485746.3) contains three GH31 modules. In addition, the analysis revealed two other human proteins (GenPeptXP_941326.2 and XP_941345.1), which are absent from the CAZy database and each contain two GH31 modules. These proteins correspond to two recently sequenced mRNAs; their genes are on chromosome 7.

A detailed nucleotide sequence analysis of the maltase–glucoamylase gene region from human chromosome 7 revealed five tandem DNA repeats coding for group a GH31 modules (Fig. 1b, *3*). The first two repeats correspond to the maltase–glucoamylase gene. The second and third repeats can be expressed to yield one polypeptide (GenPeptXP_941326.2). The fourth and fifth modules can exist as individual proteins (GenPeptAAI11974.1 and BAD18495.1) or as one two-module polypeptide (GenPeptXP_941345.1). The first and fourth modules are more similar to the N-terminal GH31 module of human sucrase–isomaltase, while the second, third, and fifth modules are more similar to the C-terminal module of this enzyme (Fig. 1b; *1, 3*). The second and third modules are near identical. A similar five-module structure is on chimpanzee chromosome 7 and, probably, *Macaca mulata* chromosome 3 (only fragmentary nucleotide sequences of this region are available). Mouse chromosome 6 and rat chromosome 4 contain, respectively, six and seven tandem repeats coding for GH31 modules (Fig. 1b; *4, 5*). On dog *Canis familiaris* chromosome 16 and bovine *Bos taurus* chromosome 4, the corresponding region harbors two genes, each coding for a two-module protein (Fig. 1b, *2*). The first and penultimate modules are similar to the N-terminal GH31 module of sucrase–isomaltase and the other modules are similar to its C-terminal module in all mammals with the completely sequenced genomes (and macaque). Data on the expression and possible alternative splicing of the maltase–glucoamylase gene and its 3'-flanking homologs are fragmentary and discrepant, and the number of their protein products in different mammals is impossible to infer. For the sake of simplicity, all corresponding GH31 modules (four to seven in different mammals) are regarded below as components of one polypeptide, which is conventionally termed maltase–glucoamylase.

The following nomenclature of the GH31 modules is proposed for mammalian sucrase–isomaltase (SUIS) and maltase–glucoamylase (MGA) on the basis of the above data and phylogenetic analysis (see below). The N- and C-terminal modules of sucrase–isomaltase are designated A and B, respectively. The first and second GH31 modules of maltase–glucoamylase are designated 1A and 1B, respectively; the penultimate and ultimate modules are 2A and 2B, respectively. In the case of more than four modules, the second module is 1B1 and the modules located between the second and penultimate ones are designated 1B2, 1B3, etc. For instance, the N-terminal module of human sucrase–isomaltase is designated SUIS_HUMAN_A and the fifth module of rat maltase–glucoamylase is designated MGA_RAT_1B4 (Fig. 1b). The designations of all GH31 modules of all proteins under study are summarized in Table 2.

Multiple sequence alignment of the maltase–glucoamylase and sucrase–isomaltase GH31 modules showed that some of them are hybrid in structure. In such modules, most of the amino acid sequence corresponds to type B, while the C end (corresponding to the C-terminal half of domain D in MalA and YicI, Fig. 1a) is more similar to type A modules. These two parts (corresponding to human sucrase–isomaltase regions 937–1733 and 1734–1827, GenPeptCAA45140.1) were independently considered in phylogenetic analysis.

Phylogenetic analysis included all chordate proteins possessing at least two GH31 modules. The only 31a-subfamily two-module protein of a nonchordate animal (sea urchin *Strongylocentrotus purpuratus*, ORF1_STRPU) was used as an outgroup to root the trees. In total, 21 amino acid sequences containing 64 GH31 modules were analyzed. All of these modules belonged to the same relatively compact cluster on the phylogenetic tree of GH31 proteins (data not shown). The use of prokaryotic MalA and YicI, whose 3D structures were experimentally established [6, 7], as an outgroup was inexpedient because of their low sequence similarity to the chordate proteins. For


MGA_PANTR_1B2
MGA_HUMAN_1B2
MGA_HUMAN_1B1
MGA_PANTR_1B1
MGA_MACMU_1B2
MGA_MACMU_1B1
MGA_CANFA_1B
MGA_BOSTA_1B
MGA_RAT_1B1
MGA_RAT_1B2
MGA_MOUSE_1B1
MGA_MOUSE_1B2
MGA_RAT_1B3
MGA_RAT_1B4
MGA_MOUSE_1B3
II
MGA_HUMAN_2B
MGA_PANTR_2B
MGA_MACMU_2B
MGA_CANFA_2B
MGA_BOSTA_2B
MGA_MOUSE_2B
MGA_RAT_2B
IV
ORF1_CHICK_B
SUIS_PANTR_B
SUIS_HUMAN_B
SUIS_MACMU_B
SUIS_BOSTA_B
SUIS_CANFA_B
SUIS_RABIT_B
SUIS_RAT_B
SUIS_MOUSE_B
SUIS_SUNMU_B
VI
B
ORF1_TETNI_B
MGA_HUMAN_2A
MGA_PANTR_2A
MGA_MACMU_2A
MGA_BOSTA_2A
MGA_CANFA_2A
MGA_MOUSE_2A
MGA_RAT_2A
III
MGA_HUMAN_1A
MGA_PANTR_1A
MGA_MACMU_1A
MGA_CANFA_1A
MGA_BOSTA_1A
MGA_MOUSE_1A
MGA_RAT_1A
I
ORF1_CHICK_A
SUIS_PANTR_A
SUIS_HUMAN_A
SUIS_MACMU_A
SUIS_BOSTA_A
SUIS_RABIT_A
SUIS_CANFA_A
SUIS_SUNMU_A
SUIS_RAT_A
SUIS_MOUSE_A
V
A
ORF1_TETNI_A
ORF1_CIOIN_B
ORF1_CIOSA_B
ORF1_CIOIN_A
ORF1_CIOIN_A
ORF1_STRPU_B
ORF1_STRPU_A
1000

instance, module SUIS_HUMAN_B (region 1201–1687, GenPeptCAA45140.1) had only 23 and 20% sequence identity to the corresponding regions of MalA and YicI, respectively, even when the most differing sequence regions belonging to domains N and D (Fig. 1a) were excluded from the analysis.

The NJ (Fig. 2) and MP (not shown) phylogenetic trees of GH31 modules were similar in topology. All vertebrate modules formed two stable clusters, corresponding to module types A and B. The only exception was module ORF1_TETNI_B, which did not belong to cluster B on the MP tree but rather formed the outgroup of this cluster together with ascidian module ORF1_CIOSA_B. The instable position of ORF1_CIOSA_B on the tree is probably explained by the fact that only a short fragment of this module is available (Table 2). Pairwise sequence comparisons of ORF1_CIOSA_B with ORF1_CIOIN_B (closest neighbor on the NJ tree) and ORF1_TETNI_B (closest neighbor on the MP tree) revealed 69 and 40% sequence identity, respectively. Hence, the topology of the NJ tree more reliably reflects the evolution of GH31 modules. When module ORF1_CIOSA_B was excluded from the analysis, the resulting MP tree (data not shown) was similar in topology to the NJ tree (Fig. 2).

Clusters A and B each included three distinct subclusters (I–VI) of mammalian proteins; the bootstrap support of each subcluster was more than 99% on each tree (NJ and MP). The subclusters were numbered so that their members occurred in the order of increasing number (from I to IV) in the dog and bovine maltase–glucoamylase sequences and the N- and C-terminal modules of sucrase–isomaltase were in subclusters V and VI, respectively. The odd-numbered clusters (I, III, and V) contained only type A modules; the even-numbered clusters contained only type B modules (Fig. 2). The phylogenetic trees of the C-terminal regions of GH31 modules (Fig. 1a) similarly had six stable subclusters of mammalian proteins (Fig. 3). Although the sequences under study were far shorter in this case, five of the six subclusters had a bootstrap support of at least 93% on both trees; only subcluster II had 90.5 and 77.5% support on the NJ and MP trees, respectively. Some sequences occurred in subcluster I rather than in subcluster II as on the trees obtained with the remaining part of the module sequences (Fig. 2). This finding confirmed the hybrid character of some modules in proteins containing more than four modules (Fig. 1b). In addition, phylogenetic trees were constructed on the basis of multiple sequence alignment of two-module structures: sucrase–isomaltase and two two-module fragments of maltase–glucoamylase (Fig. 4). The maltase–glucoamylase fragments included the first N-terminal and the third C-terminal modules (subclusters I and II) and the two last modules (subclusters III and IV). These four maltase–glucoamylase modules did not have a hybrid structure in any mammal (Fig. 1b). Clusters corresponding to maltase–glucoamylase and sucrase–isomaltase were distinctly separate (more than 98% bootstrap support) on the NJ and MP trees (Fig. 4). The first cluster included distinct subclusters corresponding to each of the two-module parts of maltase–glucoamylase (I/II and III/IV, at least 99% support on both trees).

DISCUSSION

Screening of the genome sequences of six mammalian species (cattle, rat, mouse, dog, human, and chimpanzee) revealed the maltase–glucoamylase and sucrase–isomaltase genes, which were on different chromosomes in each species. Pairwise sequence comparisons of the amino acid sequences and phylogenetic analysis make it possible to clearly distinguish the two mammalian paralogous enzymes. The two genes result from a tandem duplication of an ancestral gene (Fig. 5): sucrase–isomaltase always consists of two homologous GH31 modules, while the number of modules in maltase–glucoamylase varies. There are four modules in the simplest case (as in the bovine and dog enzymes), suggesting two consecutive tandem duplications of the full-length gene. The primate (human, chimpanzee, and macaque) enzymes have five GH31 modules; i.e., a third tandem duplication involved only a part of the ancestral gene. Duplications were probably even more numerous in rodents and led to the appearance of six- or seven-module maltase–glucoamylase gene in mouse and rat, respectively. All additional GH31 modules (located between the first N-terminal and the third C-terminal ones) in primates and rodents are always hybrid (Fig. 1b, *3–5*). This indicates that the third and further duplications did not involve a gene region coding for a full-size GH31 module, but rather a region coding for the C-terminal part of one (type A) module and most of

Fig. 2. Phylogenetic tree of maltases–glucoamylases and sucrases–isomaltases. The tree was constructed by the NJ method, using the PHYLIP software package (http://evolution.gs.washington.edu/phylip.html). Construction was based on the multiple sequence alignment of GH31 modules without the C-terminal regions (see text and Fig. 1). Statistical reliability was estimated by bootstrap analysis; bootstrap support is indicated for each node as the number of supporting pseudoreplicates out of the total of 1000. The modules are designated as in Table 2. Two-module protein ORF1_STRPU of sea urchin *Strongylocentrotus purpuratus* was used as an outgroup. A and B are two clusters of vertebrate GH31 modules; the subclusters of mammalian modules are numbered with Roman numerals. Subclusters I–IV harbor maltase–glucoamylase modules; subclusters V and VI harbor sucrase–isomaltase modules.

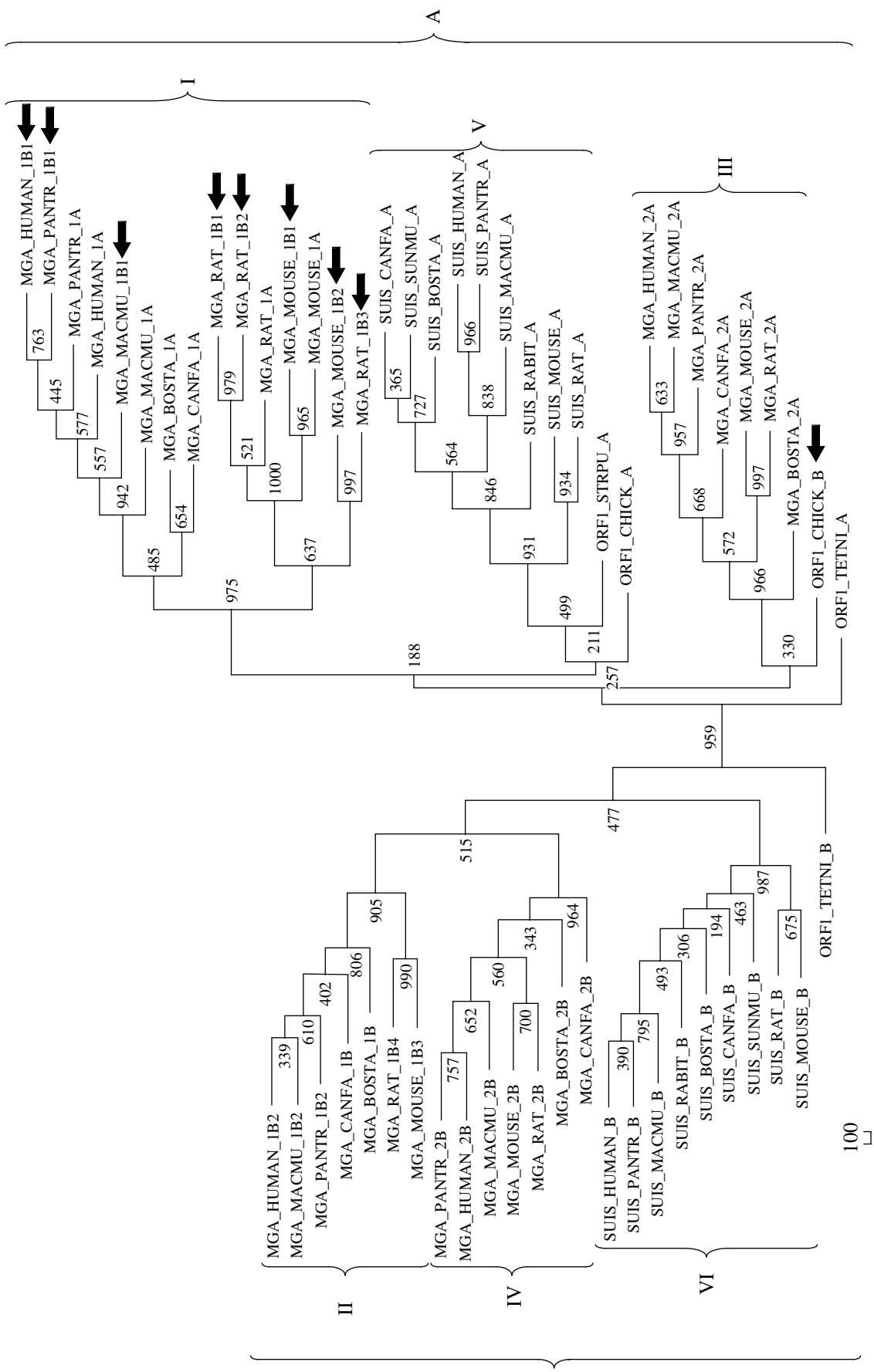


Fig. 3. Phylogenetic tree of maltases–glucoamylases and sucrases–isomaltases. The tree was constructed by the NJ method, using the multiple sequence alignment of the C-terminal module regions, corresponding to the second half of domain D (see text and Fig. 1). The cluster A modules that occur in cluster B on the previous tree (Fig. 2) are indicated with filled arrows. See Fig. 2 for other designations.

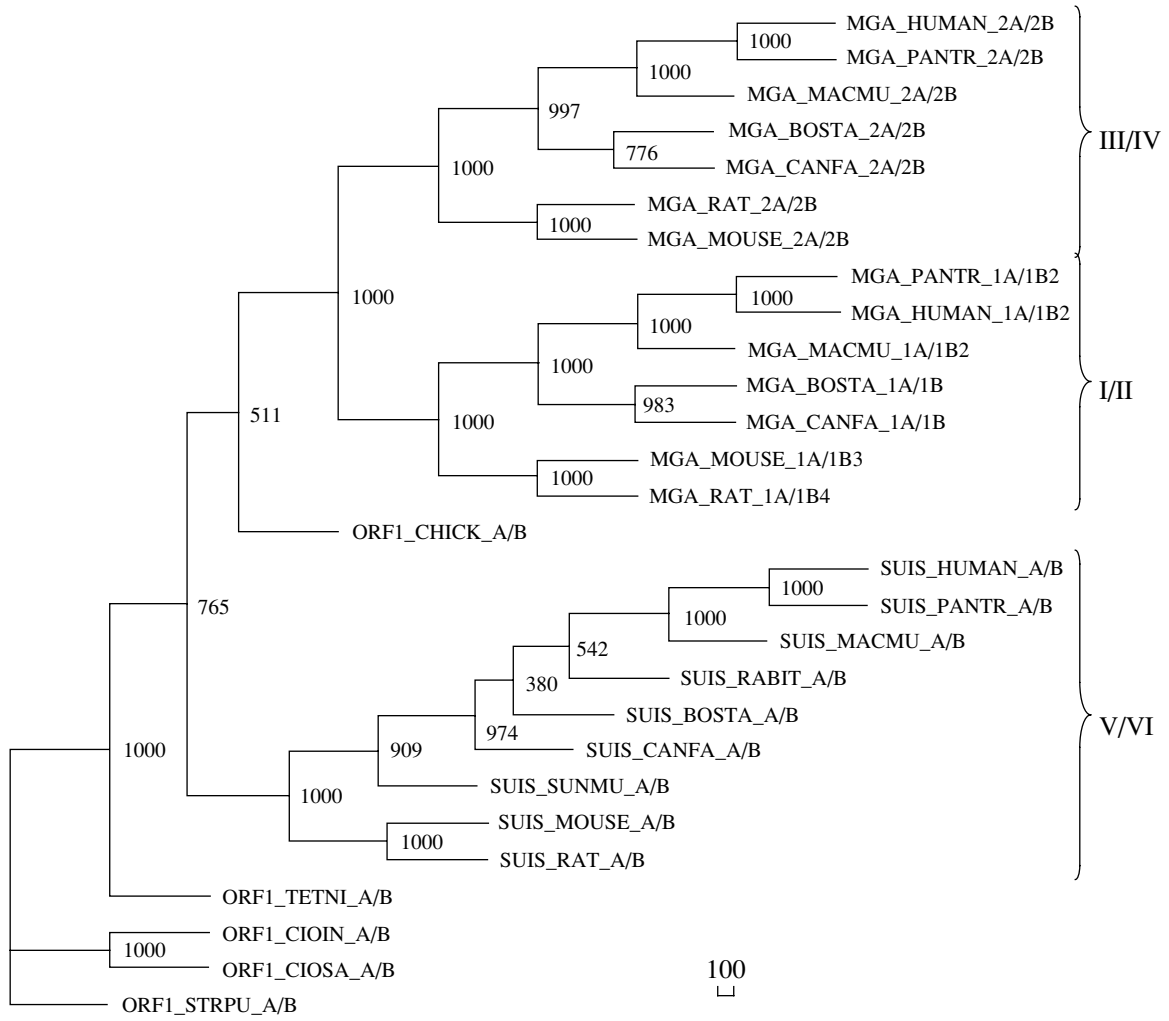


Fig. 4. Phylogenetic tree of maltases–glucoamylases and sucrases–isomaltases. The tree was constructed by the NJ method, using the multiple sequence alignment of binary GH31 modules: the two modules of sucrase–isomaltase and two two-module fragments of maltase–glucoamylase. The maltase–glucoamylase fragments included the first N-terminal and the third C-terminal modules (subclusters I and II, Fig. 2) and the two last modules (subclusters III and IV, Fig. 2). Module designations in pairs are essentially the same as in Table 2 but truncated: for instance, MGA_HUMAN_1A/1B2 stands for the first (1A) and third (1B2) modules of human maltase–glucoamylase. Sea urchin *Strongylocentrotus purpuratus* ORF1_STRPU was used as an outgroup. The subclusters are designated as in Fig. 2.

the downstream (type B) module, which together correspond to a full-size module (Fig. 1b, *6*).

Two-module homologs of the two enzymes were found in other chordates (birds, fish, and ascidia), suggesting an evolutionarily ancient origin for the two-module ancestor of maltase–glucoamylase and sucrase–isomaltase (Fig. 5). The two-module structure of sea urchin *S. purpuratus* ORF1_STRPU (echinodermata) probably results from an independent duplication. This is supported by the topology of the NJ and MP trees (Fig. 2), where the cluster consisting of ORF1_STRPU_A and ORF1_STRPU_B has a bootstrap support of more than 98%. The similarity between these modules is considerably higher than that between the GH31 modules of one protein in chordates: the two modules have 66% sequence identity in ORF1_STRPU, 49% sequence identity in ORF1_CIOIN, 35% sequence identity in ORF1_TETNI, and 43% in ORF1_CHICK. It is probable that the chordate gene for a GH31 module underwent the first tandem duplication before the advent of vertebrates (although the topology of the trees and the bootstrap support of the corresponding nodes do not exclude independent origins of the ascidian and vertebrate two-module structures). The next duplication, resulting in paralogous maltase–glucoamylase and sucrase–isomaltase, arose in vertebrates (Fig. 5). The topology of the phylogenetic trees (Figs. 2, 4) indicates that two-module ORF1_TETNI of fish *Tetraodon nigroviridis* diverged from the common ancestor before the divergence of maltase–glucoamylase and sucrase–isomaltase. Since the position of chicken

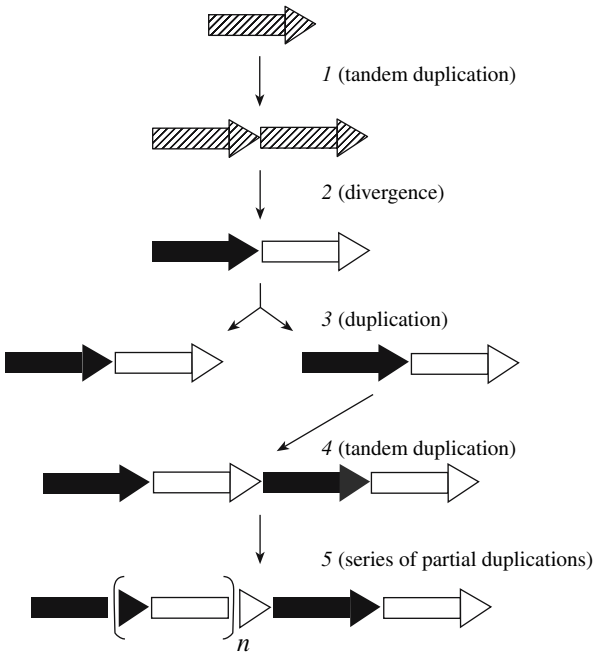


Fig. 5. Evolution of maltase–glucoamylase and sucrase–isomaltase genes. Gene regions coding for GH31 modules are shown with arrows. Filled arrows, type A modules; open arrows, type B modules; crosshatched arrows, ancestral gene (before divergence into types A and B). The N- and C-terminal regions of the modules are indicated as in Fig. 1. Evolutionary steps: *1*, tandem duplication of the ancestral gene (in primitive chordates); *2*, divergence of type A and type B modules; *3*, duplication of the two-module gene and the origin of sucrase–isomaltase and a maltase–glucoamylase precursor (probably, in primitive mammals); *4*, tandem duplication of the full-length gene and the origin of maltase–glucoamylase (which occurs in this form in cattle and dog); *5*, series of tandem duplications of the maltase–glucoamylase fragment including the C-terminal region of the first module and the N-terminal region of the second module (one duplication event in primates, two events in mouse, and three events in rat).

ORF1_CHICK is unstable on the trees, it is impossible to decide whether the last common ancestor of mammals and birds had individual genes for maltase–glucoamylase and sucrase–isomaltase or the ancestral gene was duplicated more recently (in reptiles or primitive mammals).

Apart from the proteins examined in this work, more than one GH31 domain was found in a biochemically uncharacterized protein (GenPeptXP_796510.1) of sea urchin *S. purpuratus.* The protein contains three GH31 domains corresponding to subfamily 31c. It is clear that its gene results from a series of tandem duplications, independent from duplications in the 31a-subfamily genes.

The presence of two or more domains of one family is similarly unusual for proteins possessing catalytic domains of other glycoside hydrolase families. Only a few such proteins are known. Some microorganisms (bacteria, protists, and fungi) have proteins with two to four GH5 domains [25–27]. In all cases, the domains of one protein are more similar to each other (usually having more than 90% sequence identity) than to GH5 domains of any other protein. This indicates that the genes for each of these proteins independently arose from one-module genes as a result of one or several tandem duplications. Two protist *Trichomonas vaginalis* proteins (GenPeptEAX91595.1 and EAY22311.1) each possess two GH36 domains. Pairwise sequence comparisons of their domains (which all belong to the 36D subfamily according to the classification proposed earlier [13]) indicate that their genes result from a duplication of an ancestral gene coding for a two-module protein. Proteins harboring two GH13 domains each have been found in five Firmicutes. In all cases, the domains from one protein had less than 30% similarity, ruling out tandem duplication in the origin of their genes. Pairwise sequence comparisons of these proteins suggests that the genes for four proteins—putative amylases-pullulanases [EC 3.2.1.1 and 3.2.1.41] [28] of two *Bacillus* sp. strains (GenPeptBAA11332.1 and EAR68733.1) and two *Streptococcus suis* strains (GenPeptABP93225.1 and EAP40544.1)—originate from a common ancestral gene coding for a two-module protein and that the *Clostridium perfringens* gene (GenPeptABG83674.1) results from an independent fusion of two distantly related genes. It should be noted that domains of the GH5, GH13, and GH36 families have the $(\beta/\alpha)_8$-barrel structure, like GH31 domains. Bifunctional β-xylosidase–arabinofuranosidase [EC 3.2.1.37 and 3.2.1.55] of bacterium *Caldicellulosiruptor saccharolyticus* contains two catalytic domains of the GH43 family [22]. The domains are only distantly related to each other and belong to different subfamilies (43c and 43d); i.e., the gene for this protein could not result from tandem duplication of an ancestral gene. The gene probably originates via fusion of two highly divergent homologous genes, which were juxtaposed in a *C. saccharolyticus* ancestor. Thus, chordate maltases–glucoamylases and sucrases–isomaltases are a unique group of glycoside hydrolases that evolved as proteins consisting of two (or more) modules for a long time (Fig. 5), while other two-module glycoside hydrolases arose relatively recently and, frequently, independently from each other. Further investigations are necessary to clarify the causes of the evolutionary stability of the multimodular structure in sucrase–isomaltase and, especially, maltase–glucoamylase. Yet the most probable cause is that animals need multifunctional enzymatic machinery to utilize a complex polysaccharide, starch. This assumption is supported by the fact that the most intricate maltase–glucoamylase structure is characteristic of rodents, whose diet mostly includes plant foods.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project no. 06-04-49079-a) and Young Researcher grants (MK-118.2003.04 and MK-1461.2005.4) from the President of the Russian Federation.

REFERENCES

1. Hunziker W., Spiess M., Semenza G., Lodish H.F. 1986. The sucrase–isomaltase complex: Primary structure, membrane orientation, and evolution of a stalked, intrinsic brush border protein. *Cell*. **46**, 227–234.
2. Nichols B.L., Eldering J., Avery S., Hahn D., Quaroni A., Sterchi E. 1998. Human small intestinal maltase–glucoamylase cDNA cloning. Homology to sucrase–isomaltase. *J. Biol. Chem*. **273**, 3076–3081.
3. Nichols B.L., Avery S., Sen P., Swallow D.M., Hahn D., Sterchi E. 2003. The maltase–glucoamylase gene: Common ancestry to sucrase–isomaltase with complementary starch digestion activities. *Proc. Natl. Acad. Sci. USA*. **100**, 1432–1437.
4. Henrissat B. 1991. A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J*. **280**, 309–316.
5. Carbohydrate-Active Enzymes server. 2007 (http://www.cazy.org).
6. Lovering A.L., Lee S.S., Kim Y.W., Withers S.G., Strynadka N.C. 2005. Mechanistic and structural analysis of a family 31 α-glycosidase and its glycosyl-enzyme intermediate. *J. Biol. Chem*. **280**, 2105–2115.
7. Ernst H.A., Leggio L.L., Willemoes M., Leonard G., Blum P., Larsen S. 2006. Structure of the *Sulfolobus solfataricus* α-glucosidase: Implications for domain conservation and substrate recognition in GH31. *J. Mol. Biol*. **358**, 1106–1124.
8. Naumoff D.G. 2001. Sequence analysis of glycosylhydrolases: β-fructosidase and α-galactosidase superfamilies. *Glycoconjugate J*. **18**, 109.
9. Naumoff D.G. 2002. Sequence analysis and classification of α-galactosidases. *International Summer School "From Genome to Life: Structural, Functional and Evolutionary Approaches,"* Cargèse, Corsica, France, p. 40 (http://www-archbac.u-psud.fr/Meetings/cargese2002/).
10. Rigden D.J. 2002. Iterative database searches demonstrate that glycoside hydrolase families 27, 31, 36 and 66 share a common evolutionary origin with family 13. *FEBS Lett*. **523**, 17–22.
11. Naumoff D.G. 2003. α-Galactosidase superfamily: Phylogenetic analysis and homology with some α-glucosidases. *Program and Abstracts of the 5th Carbohydrate Bioengineering Meeting*, Groningen, the Netherlands, p. 81.
12. Naumoff D.G. 2004. Phylogenetic analysis of α-galactosodases of the GH27 family. *Mol. Biol*. **38**, 463–467.
13. Naumoff D.G. 2004. The α-galactosidase superfamily: Sequence based classification of α-galactosidases and related glycosidases. *Proc. Fourth Int. Conf. on Bioinformatics of Genome Regulation and Structure*, July 25–30, 2004, Novosibirsk, Russia, vol. 1, pp. 315–318 (http://www.bionet.nsc.ru/meeting/bgrs2004/tom1.pdf).
14. Naumoff D.G. 2006. Development of a hierarchical classification of the TIM-barrel type glycoside hydrolases. *Proc. Fifth Int. Conf. on Bioinformatics of Genome Regulation and Structure*, July 16–22, 2006, Novosibirsk, Russia, vol. 1, pp. 294–298 (http://www.bionet.nsc.ru/meeting/bgrs2006/BGRS_2006_V1.pdf).
15. McCarter J.D., Withers S.G. 1996. Unequivocal identification of Asp-214 as the catalytic nucleophile of *Saccharomyces cerevisiae* α-glucosidase using 5-fluoro glycosyl fluorides. *J. Biol. Chem*. **271**, 6889–6894.
16. Henrissat B. 1998. Glycosidase families. *Biochem. Soc. Trans*. **26**, 153–156.
17. Ernst H.A., Leggio L.L., Yu S., Finnie C., Svensson B., Larsen S. 2005. Probing the structure of glucan lyases by sequence analysis, circular dichroism and proteolysis. *Biologia* (Bratislava). **60**, 149–159.
18. Naumoff D.G. 2005. GH97 is a new family of glycoside hydrolases, which is related to the α-galactosidase superfamily. *BMC Genomics*. **6**, Art. 112.
19. Janeček Š., Svensson B., Macgregor E.A. 2007. A remote but significant sequence homology between glycoside hydrolase clan GH-H and family GH31. *FEBS Lett*. **581**, 1261–1268.
20. Naumoff D.G. 2006. Gene structure and evolution of mammalian maltase–glucoamylase. *Abstr. XXIII Int. Carbohydrate Symposium*, July 23–28, 2006, Whistler, Canada, p. 244. Abstr. THU-C4-AM.1.
21. Naumoff D.G., Livshits V.A. 2001. Molecular structure of the *Lactobacillus plantarum* sucrose utilization locus: Comparison with *Pediococcus pentosaceus*. *Mol. Biol.* **35**, 19–27.
22. Naumoff D.G. 2001. β-Fructosidase superfamily: Homology with some α-L-arabinases and β-D-xylosidases. *Prot. Struct. Funct. Genet*. **42**, 66–76.
23. Kuznetsova A.Y., Naumoff D.G. 2006. Phylogenetic analysis of COG1649, a new family of predicted glycosyl hydrolases. *Proc. Fifth Int. Conf. on Bioinformatics of Genome Regulation and Structure*, July 16–22, 2006, Novosibirsk, Russia, vol. 3, pp. 179–182 (http://www.bionet.nsc.ru/meeting/bgrs2006/BGRS_2006_V3.pdf).
24. Naumoff D.G. 2006. Phylogenetic analysis of a protein family. *Zbio*. **1**, Art. 3 (http://zbio.net/bio/001/003.html).
25. Aylward J.H., Gobius K.S., Xue G.-P., Simpson G.D., Dalrymple B.P. 1999. The *Neocallimastix patriciarum* cellulase, CelD, contains three almost identical catalytic domains with high specific activities on Avicel. *Enzyme Microb. Technol*. **24**, 609–614.
26. Eberhardt R.Y., Gilbert H.J., Hazlewood G.P. 2000. Primary sequence and enzymic properties of two modular endoglucanases, Cel5A and Cel45A, from the anaerobic fungus *Piromyces equi*. *Microbiology.* **146**, 1999–2008.
27. Yoda K., Toyoda A., Mukoyama Y., Nakamura Y., Minato H. 2005. Cloning, sequencing, and expression of a *Eubacterium cellulosolvens* 5 gene encoding an endoglucanase (Cel5A) with novel carbohydrate-binding modules, and properties of Cel5A. *Appl. Environ. Microbiol*. **71**, 5787–5793.
28. Hatada Y., Igarashi K., Ozaki K., Ara K., Hitomi J., Kobayashi T., Kawai S., Watabe T., Ito S. 1996. Amino acid sequence and molecular structure of an alkaline amylopullulanase from *Bacillus* that hydrolyzes α-1,4 and α-1,6 linkages in polysaccharides at different active sites. *J. Biol. Chem*. **271**, 24,075–24,083.