The maltase-glucoamylase gene: Common ancestry to sucrase-isomaltase with complementary starch digestion activities

Buford L. Nichols*†, Stephen Avery*, Partha Sen*, Dallas M. Swallow‡, Dagmar Hahn§, and Erwin Sterchi§

*U.S. Department of Agriculture, Agricultural Research Service, Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030-2600; ‡Galton Laboratory, Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, United Kingdom; and §Institute of Biochemistry and Molecular Biology, University of Bern, Buehlstrasse 28, CH-3012 Bern, Switzerland

Communicated by John Waterlow, University of London, London, United Kingdom, December 4, 2002 (received for review April 17, 2002)

Brush-border maltase-glucoamylase (MGA) activity serves as the final step of small intestinal digestion of linear regions of dietary starch to glucose. Brush-border sucrase-isomaltase (SI) activity is complementary, through digestion of branched starch linkages. Here we report the cloning and sequencing of human MGA gene and demonstrate its close evolutionary relationship to SI. The gene is ≈82,000 bp long and located at chromosome 7q34. Forty-eight exons were identified. The 5′ gene product, when expressed as the N-terminal protein sequence, hydrolyzes maltose and starch, but not sucrose, and is thus distinct from SI. The catalytic residue was identified by mutation of an aspartic acid and was found to be identical with that described for SI. The exon structures of MGA and SI were identical. This homology of genomic structure is even more impressive than the previously reported 59% amino acid sequence identity. The shared exon structures and peptide domains, including proton donors, suggest that MGA and SI evolved by duplication of an ancestral gene, which itself had already undergone tandem gene duplication. The complementary human enzyme activities allow digestion of the starches of plant origin that make up two-thirds of most diets.

family 31 glucoside hydrolases | small intestine | granulocyte

The main digestible carbohydrates in the human diet are starch and sucrose. Plant starches provide the largest percentage of calories in the diet, and sucrose, the precursor of starch synthesis, is a major contributor to the esthetic quality of the diet (1). Mucosal maltase-glucoamylase (MGA; EC 3.2.1.20 and 3.2.1.3, encoded by a gene, *MGAM*, located on chromosome 7) activity serves as the final step in small-intestinal digestion of linear regions of starch to glucose (2). Mucosal sucrase-isomaltase (SI) (EC 3.2.148 and 3.2.1.10, encoded by a gene, *SI*, on chromosome 3q26) activity constitutes the last stage of small-intestinal digestion of branch points of starch to glucose. Thus, these two enzymes complement one another in the digestion of starch. In a recent paper, we reported cloning and sequencing of human small-intestinal MGA cDNA (ref. 2; GenBank accession no. NM_004668). MGA has two catalytic sites, which are identical to those of SI. MGA and SI are members of glycosyl hydrolase family 31, but the proteins show only 59% amino acid sequence identity and have complementary activities in plant carbohydrate digestion (2).

Starch granules are a mixture of two different plant polysaccharides, amylose, a linear [4-*O*-α-D-glucopyranosyl-D-glucose]$_n$ polymer, and amylopectin, with additional 6-*O*-α-D-glucopyranosyl-D-glucose links (≈4% of total), which result in a branched structure. Dietary starches are a mixture of ≈25% amylose with 75% amylopectin, a fact of nutritional significance because of the complexity of the mammalian starch digestion pathway (3). In the small intestine, α-amylase (1,4-α-D-glucan glucohydrolase, EC 3.2.1.1, chromosome 1p21) is an endohydrolase, found in mature salivary and pancreatic secretions, that produces solubilized linear maltose oligosaccharides by hydrolysis of internal α1-4 linkages (3, 4). α-Amylase bypasses the α1-6 linkages of amylopectin and thus produces branched maltose oligosaccharides. Both families of maltose oligosaccharides are not absorbable without further processing to glucose by hydrolysis at the nonreducing ends of 1-4 and 1-6 oligomers (3). Two different mammalian small intestinal mucosal brush border-anchored enzymes, MGA and SI, carry out this hydrolysis to glucose (3). Enzyme substrate specificities of human SI complement those of MGA. *In vivo*, SI accounts for 80% of neutral mucosal maltase (1,4-*O*-α-D-glucanohydrolase) activity, all neutral sucrase (D-glucopyranosyl-β-D-fructohydrolase) activity, and almost all isomaltase (1,6-*O*-α-D-glucanohydrolase) activity (3, 5, 6). MGA accounts for all mucosal neutral glucoamylase exoenzyme (1,4-*O*-α-D-glucanohydrolase) activity for amylose and amylopectin substrates, 1% of isomaltase activity, and 20% of neutral maltase activity (3, 5, 6). The spectrum of MGA and SI activities, with complementary substrate specificities, is thus indispensable for small-intestinal digestion of the plant-derived α-D-glucose oligomers to glucose (3).

Here we report the sequencing and mapping of the gene structure of *MGAM* and provide further evidence for (*i*) a common ancestry for MGA and SI, the enzymes essential for small-intestinal digestion of plant α-D-glycosides to glucose, (*ii*) the enzyme activity and substrate specificity of the recombinant protein expressed by 5′ MGA cDNA, and (*iii*) the homology of genomic architecture and peptide domains of MGA with its complementary enzyme SI.

Methods

Ethical Aspects. The ethical committees of Baylor College of Medicine and related institutions (2, 7) approved this investigation.

Cloning Procedures. PCR was carried out as described (2). Amplicons were separated, purified, and transformed into NovaBlue *Escherichia coli* (Novagen). Clones were screened by PCR and plasmid isolations (Qiagen, Valencia, CA) were carried out. DNA was sequenced (373A automated sequencer, Applied Biosystems) at Baylor College of Medicine Child Health Research Center Core Laboratory. When exon-specific primer pairs were not available, the GenomeWalker technology (CLONTECH) was used to extend genomic sequences with cDNA-based and adaptor primers. Primary and secondary nested PCR was performed according to the manufacturer's protocol. A PAC library R81–84, provided by the Cloning Core Laboratory of the Baylor College of Medicine Human Genome

Abbreviations: MGA, maltase-glucoamylase; SI, sucrase-isomaltase; TFF, trefoil peptide domain; GAA, acid glucosidase alpha (a lysosomal α-glucosidase).

Data deposition: The MGAM1–MGAM21 nucleotide sequences reported in this paper have been deposited in the GenBank database (accession nos. AF432182–AF432202).

†To whom correspondence should be addressed at: Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine, 1100 Bates Street, Houston, TX 77030-2600. E-mail: bnichols@bcm.tmc.edu.

Center, was screened with MGA-specific primers and confirmed by hybridization.

Computer Analysis. The software from the Genetics Computer Group (8) was accessed via the Baylor College of Medicine Molecular Biology Computation Resource (9). Sequence analysis was performed with the Genetics Computer Group programs (8). GenBank data were searched using BLAST (10) or FASTA and TFASTA (11) programs. Primer designing was done with the PRIMER (8) program. Phylogenetic and molecular evolutionary analyses were conducted using a Neighbor Joining matrix calculation of distance of internal nodes based on divergence of peptide sequences with MEGA V.2.1 software (12), with default parameters, and sequences downloaded from the ProDom peptide domain site (13).

Additional *MGAM* clones were found by searching GenBank and Celera blasts with MGA cDNA sequences (10, 11). In each case there was extensive high-quality alignment (>98%) between MGA and the GenBank and Celera clone sequences. A genomic structure of *SI* was available by searching, with SI cDNA sequences, the 5′ end (14) and the 3′ end from GenBank (contig AC021100) and Celera (transcript hCT31261/gene hCG40008; scaffold GA_2HTBL5CBJJ; and unassembled fragment files). All MGA and SI exon boundary locations were confirmed by cDNA sequences and the GRAIL 3 exon identification program (http://grail.lsd.ornl.gov/grailexp/). Peptide folds and predicted tertiary structures were classified with 3D-PSSM (15).

Recombinant Expression. The 5′ domain of MGA cDNA was isolated from clone MGA-P1$_2$ and subcloned into pSGKS (16) to yield clone MGA-P1A. The cDNA after the single *Pml*I restriction site (nc603) was substituted by a 2260-bp *Pml*I/*Not*I fragment isolated from MGA-P1A$_2$ (2) encoding residues 202–954 yielding clone MGA-P1A$_2$. The orientation was confirmed by sequencing. A D529A mutation was produced by recombinant PCR (16). The mutated product was ligated back into the MGA-P1A$_2$ vector and sequenced. COS-1 cells were grown and transfected as described (7, 17). Cells were pulse-labeled with 50 μCi (1 Ci = 37 GBq) [^{35}S]methionine (NEN). The recombinant proteins were immunoisolated with monoclonal antiserum HBB 2/143/17 and treated with endo H (endo-*N*-acetylglucosaminidase H) before analysis by SDS/PAGE as described (7, 17). Transfected cells were isolated and homogenized as described (17). Fifty microliters of homogenate was assayed for soluble starch (amylose) and 25 μl for maltose and sucrose (2% substrate) hydrolysis to glucose under standard conditions used for human enzyme assay (7). Glucose was quantified by glucose oxidase (7). Protein was measured using bicinchoninic acid (ref. 18; Pierce).

Results

Processing and Enzyme Activity of Recombinant MGA. The COS cell expression of the MGA-P1A$_2$ demonstrated synthesis and processing of isoforms of MGA as in organ culture experiments (7). The high-mannose and complex glycosylated isoforms of P1A$_2$ were identified by endo H enzyme sensitivity (ref. 7; Fig. 1). The sequential glycosylation of the P1A$_2$ high-mannose and complex glycosylated forms was also demonstrated by pulse labeling (not shown). Enzymatic activity of P1A$_2$ recombinant protein was present with maltose and soluble starch (amylose) substrates, but negligible activity was present with lactose or sucrose substrates (Table 1). Mutation of the D to A in the signature I site, WIDMNE, resulted in loss of P1A$_2$ activity for maltose or starch hydrolysis (Table 1).

Cloning and Sequencing of Genomic *MGAM*. An initial PAC clone, containing *MGAM*, was identified from library STS M471/M604

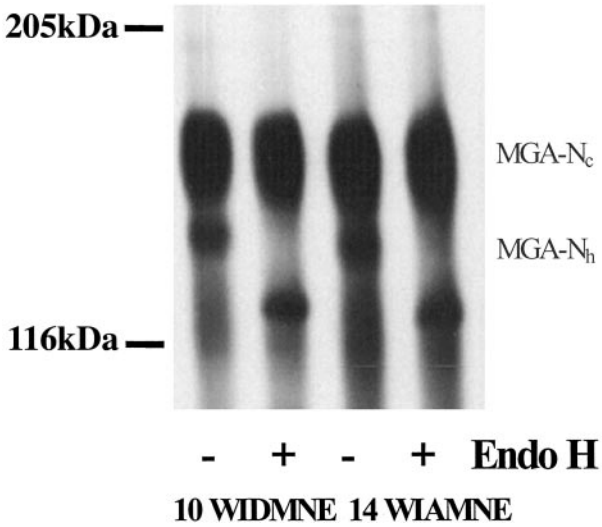

Fig. 1. Expression and processing of the wild-type and D529A-mutated N-terminal region of MGA in a pulse-labeling experiment. The high-mannose-specific enzyme endo H (+) was used to identify the isoforms; note the reductions in the size of the high-mannose isoforms and the stability of complex glycosylated isoforms. The high-mannose isoforms are indicated as MGA-N$_h$ and the complex glycosylated forms as MGA-N$_c$. Mutation of the aspartic acid proton donor, in clone 14, to alanine (D529A) resulted in loss of all enzymatic activity (Table 1) without altering cellular processing from that of wild-type clone 10.

and filter 81–84. It was identified as PAC clone 81E19 (Fig. 2). Using cDNA primers, it was determined that the clone extended 5′ from exon 33 through intestinal exon 1. The cDNA sequence also was used to search for additional *MGAM* genomic clones in GenBank. A BAC clone R-1083c11, from library RPCI-11, was identified (AQ743657) that matched MGA (98%) at the 3′ end from cDNA +4112 to +4172 and extended beyond the poly(A) tail (Fig. 2). The sequenced 3′ end of clone R-1083c11 includes AQ744869, which matches (97%) four exons and introns of the human germline T cell receptor beta chain gene (contained in contig U66059, U69054 to U68742), mapped to 7q35. Using publicly available overlapping sequences (accession nos. AC011654 and AC073647), a long chromosome 7 contig (NT_023529) extending 5′ from *MGAM* exon 10 was constructed. This extended the 5′ *MGAM* sequence to −20147 bases upstream of the MGA intestinal exon 1, confirmed all PAC 81E19 sequences through exon 10 (99% match), and included the granulocyte MGA exon 1 (unpublished cDNA). After submission of our *MGAM* sequences to GenBank, an additional chromosome 7 contig (NT_023640) was reported. This new unassembled contig includes *MGAM* exons 1–48. The 5′ end of PAC clone 81E19 began +18650 bp from the AC011654 contig 3′ end, and the 5′ end of BAC clone 2783F23 began +22323 bp from the AC073647 contig 3′ end. Two additional BAC unassembled sequences (AC091742 and AC091684) encompassed portions of

Table 1. Substrate activities (EU/g protein) of recombinant MGA-P1A$_2$ proteins resulting from clones transfected into COS-1 cells

Clone	*n*	Maltose	Starch	Sucrose	Lactose
10 (WIDMNE)	5	141.4 ± 6.7	9.2 ± 0.2	0.5 ± 0.4	0
14 (WIAMNE)	3	0	0	0	0
None (blank)	3	2.79 ± 0.07	1.04 ± 0.07	0	0

The D of the WIDMNE domain was mutated to an A. No clone was in blank vector.

PHYSIOLOGY

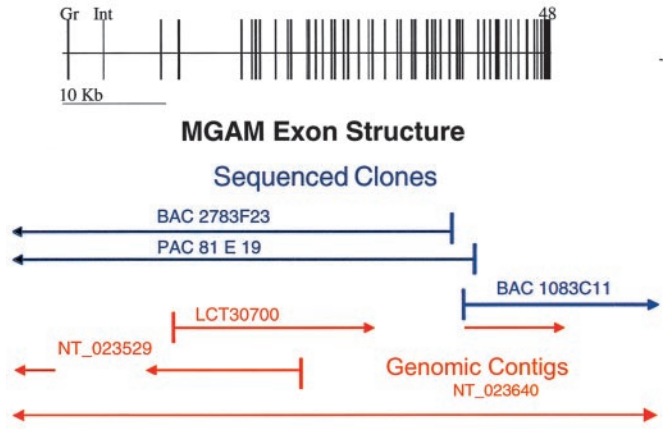

Fig. 2. Map of *MGAM* exon structure as it relates to the sequenced clones in red and the relationships to contigs in GenBank in blue. Both granulocyte (Gr) and small intestinal (Int) exons 1 were identified; these are located at the 5′ end of the gene. The 3′ end is identified as the poly(A) tail of the cDNA. The locations of exons and introns are shown to scale and a 10-kb benchmark is provided.

the *MGAM* coding region excluding granulocyte exon 1 but including intestinal exons 1 and 48. In total, 21 fragments of *MGAM* were sequenced for a total of 46,872 bases (see Fig. 4, which is published as supporting information on the PNAS web site, www.pnas.org). All five GenBank sequences and a partial genomic sequence from Celera (transcript hCT30700/gene hCG39449), which extended from exons 3 to 21 and 37 to 42, confirmed our *MGAM* sequences (>96% identity), extended the full size of the gene to ≈82,000 bases, and confirmed our exon boundaries.

Chromosomal Location. In 1998, an EST, GS1365, on chromosome 7 was identified with sequence identity to the 3′ end of MGA cDNA. This led to the suggestion that *MGAM* is located on chromosome 7 (3). The 3′ end of BAC clone 2783F23 terminates in sequence B99061, which matches MGA cDNA sequence from 4400 to 4538 (96%) and extends in the 5′ direction. The 5′ sequenced end of BAC 2783F23 ends at B99063, which matches (98%) four exons of myeloid DAP12-associating lectin (*MDL-1*), the C-type, calcium-dependent carbohydrate-recognition domain, and lectin superfamily member 5 (CLECSF5, AJ271684) gene, which has been cytogenetically mapped to 7q33. The 3′ end of the contig resulting from AC011654 and AC073647 contained 16,766 bp of the *MDL-1* gene (99% cDNA match), including seven exons and six introns, beginning at −52771. BAC clone R-1083c11, which matched MGA (98%) at the 3′ end from cDNA 4112 to 4172 and extended beyond the poly(A) tail, was identified (AQ743657). The sequenced 3′ end of clone R-1083c11 includes AQ744869, which matches (97%) four exons and introns of the human germline T cell receptor beta chain gene (*TCRB*, contained in contig U66059, 69054–68742), mapped to 7q35. These two BAC clones with sequenced ends establish that *MGAM* is anchored 3′ of *MDL-1* at 7q33 and 5′ of *TCRB* at 7q35. These results were confirmed by the unassembled contig NT_023640, which also contains *MDL-1* and *TCRB*.

Exon Boundaries. Exon boundaries were mapped by PCR with MGA cDNA sequence-designed primers by using the PAC 81E19 and BAC 1083c11 genomic clones as templates. In the case of longer introns and the promoter, the GenomeWalker system, which used MGA cDNA sequence nested primers, was used to extend genomic sequences. Our sequences were checked against chromosome 7 BAC clones, NT_023529, NT_023640, and Celera transcript hCT30700. Forty-eight exons were identified, all with classic intron 5′ splice donor and 3′ splice acceptor sequences (Table 2). The exon sizes ranged from 35 to 201 bases, with the exception of exon 48, which had 953 bases. The location and boundary codon phasing of all MGA exons were identical to those of SI. The size of the exons was conserved between the two genes in 43 of the 48 exons but varied in the remainder (see pile-up in Fig. 5, which is published as supporting information on the PNAS web site).

Discussion

Recombinant MGA Enzyme Activity. The recombinant N-terminal domain of MGA hydrolyzed glucose from maltose and linear starch but not from sucrose or lactose substrates (Table 1). This establishes that the recombinant MGA N-terminal domain has the substrate specificity of native MGA and is distinctive from SI, confirming their complementary substrate specificities. Both enzymes hydrolyze maltose, but MGA has specificity for the 95% α1-4 and SI for the 5% α1-6 glucose linkages of starch (3, 4). The function of the MGA C-terminal site is currently under investigation. The collaboration of these enzymes in maltose clearance from the lumen may be of importance because maltose is a noncompetitive inhibitor of α-amylase, and the rate of maltose brush-border digestion could regulate luminal α-amylase endoglycosidase activity (19).

The aspartic acid, WIDMNE, at conserved site III is known to be catalytic in SI and four other family 31 enzymes (20–22). This putative MGA proton donor was mutated from D to A, and resulted in the loss of all recombinant enzyme activities. This is proof that this amino acid serves as a catalytic acid in MGA as well as SI. Studies in *Schizosaccharomyces pombe*, another family 31 α-glucosidase, revealed that mutation of either the conserved site III D or the site IV D (see pile-up in Fig. 5) reduced enzyme activity (23). This suggests that the two conserved D in sites III and IV serve as proton donors and recipients for family 31 α-glucosidases including MGA and SI.

***MGAM* and *SI* Gene Structures.** The *MGAM* gene is ≈82,000 bp long. Forty-eight exons were identified, all with classic intronic 5′ splice donor and 3′ splice acceptor sequences (Figs. 4 and 5, Table 2). At the time of submission, exon boundaries, assigned by computer to contig NT_028590, agreed with assignments in this manuscript for exon boundaries and phases 17–31 but were absent or in disagreement for the remaining 24 exons. The exon sizes ranged from 35 to 201 bases, with the exception of exon 48 (953 bases), and the whole structure represents an ancestral duplication, as surmised (2), with a very high degree of conservation of exon structure. There were 25 exons coding the N-terminal part and 23 coding the C-terminal part (Table 2). The N-terminal expression construct MGA-P1A$_2$ terminates in the middle of exon 25. Exons 1 and 2 were unique to the N-terminal part, and an exon boundary between exons 34 and 35 was unique to the C-terminal part. The exon boundary between exons 25 and 26 in the C-terminal part was extended by 10 aa from that of exons 3 and 4. A pile-up of both terminals documents a conservation of all of the other exon boundaries in both MGA domains (see Fig. 5).

The exonic structure of *SI*, as constructed from publicly available sequences, is also shown in Table 2. *SI* also appears to be ≈82,000 bp long and to have 48 exons and the same differences in boundaries between the N- and C-terminal parts as *MGAM*. Classic intron 5′ splice donor and 3′ splice acceptor sequences were identified in SI, and exon boundary phases were identical to those in MGA (Table 2). Five SI exons were shorter than those of MGA but the remaining exons were of identical sizes. Exons 1 and 48 are noncoding or partially coding and shorter than in MGA. Exons 2 and 3 code unduplicated cytoplasmic tails, membrane anchors, and glycosylated stalk regions,

Table 2. Comparison of exon sizes and boundary codon phases in MGA and SI

Exon boundary	MGA exon length, bp	SI exon length, bp	Exon size difference, bp	MGA phase	SI phase	Exon domain
1	52	62	10	—	—	3′ UTR
2	129	118	11 (9 coding)	0	0	Membrane
3	200	137	62	1	1	Stalk
4	121	118	3	0	0	TFF
5	110	110	0	1	1	β-sheet
6	152	152	0	0	0	β-sheet
7	172	172	0	2	2	β-sheet
8	100	100	0	0	0	β-sheet
9	113	113	0	1	1	$\beta\alpha$-barrel
10	126	126	0	0	0	$\beta\alpha$-barrel
11	132	132	0	0	0	$\beta\alpha$-barrel
12	117	120	3	0	0	$\beta\alpha$-barrel
13	114	114	0	0	0	$\beta\alpha$-barrel
14	85	85	0	0	0	H-donor
15	118	118	0	1	1	$\beta\alpha$-barrel
16	172	172	0	2	2	H-acceptor
17	117	117	0	0	0	Sig II
18	155	155	0	0	0	$\beta\alpha$-barrel
19	85	85	0	2	2	$\beta\alpha$-barrel
20	57	57	0	0	0	β-sheet
21	125	125	0	0	0	β-sheet
22	89	89	0	2	2	β-sheet
23	50	50	0	1	1	β-sheet
24	168	171	3	0	0	β-sheet
25	153	156	3	0	0	β-sheet
26	201	207	6	0	0	TFF
27	155	155	0	0	0	β-sheet
28	169	169	0	2	2	β-sheet
29	97	97	0	0	0	β-sheet
30	113	113	0	1	1	$\beta\alpha$-barrel
31	126	126	0	0	0	$\beta\alpha$-barrel
32	129	129	0	0	0	$\beta\alpha$-barrel
33	111	111	0	0	0	$\beta\alpha$-barrel
34	63	63	0	0	0	$\beta\alpha$-barrel
35	135	117	18	0	0	$\beta\alpha$-barrel
36	88	88	0	0	0	H-donor
37	139	139	0	1	1	$\beta\alpha$-barrel
38	134	134	0	2	2	H-acceptor
39	35	35	0	1	1	$\beta\alpha$-barrel
40	117	117	0	0	0	Sig II
41	149	149	0	0	0	$\beta\alpha$-barrel
42	85	85	0	2	2	$\beta\alpha$-barrel
43	57	57	0	0	0	β-sheet
44	125	125	0	0	0	β-sheet
45	89	89	0	2	2	β-sheet
46	50	50	0	1	1	β-sheet
47	171	168	3	0	0	β-sheet
48	963	544	401	0	0	5′ UTR

which are longer in MGA. MGA exon 35 codes an extra sequence, NPQNPE, inserted 10 amino acids before the C-terminal proton donor, not present in SI. The duplicated trefoil peptide domains (TFF) are coded by exons 3–4 and 25–26 in both genes. Although much is known about the function of TFF peptides (24), functions of TFF glucosidase domains remain unknown.

Interestingly, the two WIDMNE (signature I) MGA proton donor codons are split by an intron, as has been previously observed for GAA (glucosidase, alpha; acid; GenBank accession no. NP_000143, chromosome 17). This occurs before the D codon, between exons 13 and 14 and exons 36 and 37. The same D coding splits were conserved in the N-terminal (isomaltase) and C-terminal (sucrase) parts of SI (position 611 in the numbering of the pile-up in Fig. 5). Four putative D proton acceptors are located in a conserved sequence WLGDN within exons 16 and 38 of both MGA and SI (position 719 in the numbering of the pile-up in Fig. 5). The four signature II sequences are coded within exons 17 and 40. The function of these conserved signature II sequences is unknown.

The MGA stop codon in the C-terminal part is located 291 bases within the 963-base exon 48, whereas the stop codon of SI is 134 within a 544-base exon. The 401 base difference in size of exons 48 is the largest divergence between MGA and SI and is preserved in mice (unpublished work). The second largest is a 62-bp increase in length of MGA exon 3, which codes the stalk

regions of both enzymes. The third is the presence of an added 18 bp in MGA exon 35 at the end of the C-terminal barrel coding sequence. In contrast, SI has an additional 10 bp coding the 5′ UTR.

Predicted Tertiary Protein Structure. There have been no reported studies of the crystallographic structure of family 31 proteins. Reasonable approximations (>95%) were available via computational methods by alignment of peptide folds with those of known crystallographic structures. All four MGA and SI protein domains in Table 2 and the pile-up in Fig. 5 have predicted folds of a TIM $(\beta/\alpha)_8$-barrel type [named for *t*riose phosphate *iso*merase, the first described $(\beta/\alpha)_8$-barrel structure]. The full predicted tertiary structure for MGA and SI can be described as a β–$(\beta/\alpha)_8$–β–$(\beta/\alpha)_8$–β "club sandwich" because of three beta sheets enclosing the tandem duplicated barrels. The shared folds and predicted tertiary structures of MGA and SI are consistent with a common origin.

Exon Structure and Peptide Domains. The location of exon boundaries provides information about coded protein domains (25). The conserved structure of the four MGA and SI peptide domains in Table 2 and the pile-up in Fig. 5 permits several deductions. There were four helices located in the N direction and four in the C direction around each central signature I WIDMNE sequence. The first helix was on the N side of an exon boundary of all four peptides. The fourth helix is in the same exon as a β-sheet in the N half of the WIDMNE sequence. The signature I site is located at the C end of helix 4. The C end of helix 5 precedes a short β-sheet and loop containing a conserved WLGDN sequence; the D is a proton acceptor for family 31 hydrolases (20, 21). Signature II is at the C end of helix 6. Helix 7 is at the center of the signature II constant region and is located on the C side of a cysteine-bounded variable region. There are two extra coding exons in MGA-C- and SI-C-terminal peptides. The N-most inserts an extra helix between helix 3 and helix 4 in MGA-C- and SI-C-terminal peptides. Six of seven universally conserved sequences of family 31, shown in bold text and numbered I–VI in the pile-up in Fig. 5, are located within loops at the C end of a β-sheet (18–20). Conserved sequence VII, which lies within α-helix 8, is the exception. Locations of conserved sequences and catalytic acids at the C end of helices (in the pile-up in Fig. 5) are consistent with the catalytic sites residing on the C surface of the $(\beta/\alpha)_8$-barrels. The conservation of uniquely duplicated tandem peptide domains between MGA and SI is also consistent with the hypothesis of a common origin. No other duplicated glucoside hydrolases could be found in GenBank.

Phylogenetic Deductions. The shared exon structures and conservation of protein domains of MGA and SI are consistent with the notion that these genes arose by duplication after an initial internal duplication of the ancestral gene. The N-terminal regions of each gene are more closely related to each other, and likewise, the C-terminal regions to each other, than are the N- and C-terminal regions of the same gene. There is greater conservation of the available exonic structures of *MGAM* and *SI* (100%) than in peptide sequence (59%). The predicted $(\beta/\alpha)_8$ tertiary structure is highly conserved between MGA and SI proteins, as indicated by seven invariant family 31 glycoside hydrolase domains (ref. 22; bold type in the pile-up in Fig. 5). The earliest known phylogenetic association between TFF and α-glucosidases domains reported was from the worm *Caenorhabditis elegans*. Two *C. elegans* genes (g1065946, chromosome III, and g6425187, chromosome IV) have a single TFF preceding an unduplicated α-glucosidase domain. These *C. elegans* and mammalian lysosomal acid glucosidase (GAA) homologies are more similar to C-terminal domains of *MGAM* and *SI* genes. This

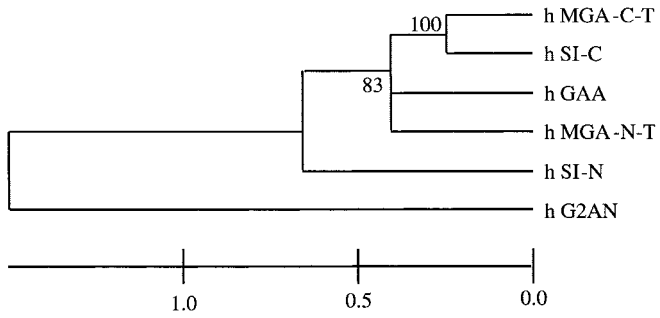

Fig. 3. Genetic relationships of the human family 31 proteins. The N- and C-terminal regions of the MGA and SI protein sequences, with duplications of signature sites, are compared by a bootstrap analysis with pair-wise deletions and Poisson correction to GAA and G2AN. The full peptide sequences are given in Fig. 5. The scale below the tree provides time relative to the present in arbitrary units. The family 31 mammalian peptide sequences are analyzed in Fig. 6.

suggests that the C-terminal part of *MGAM* and *SI* is closer to the founder of the duplicated domains. The occurrence of a single TFF-linked α-glucosidase suggests that the tandem duplications in *MGAM* and *SI* developed later in phylogeny than the *C. elegans* worm. The human TFF and $(\beta/\alpha)_8$ domain duplications are demarcations of *MGAM* and *SI* from an older unduplicated *GAA* that shares 9 of the 10 exon boundaries (*GAA* exons 6–16) within their C-terminal $(\beta/\alpha)_8$ domains, but only 2 of 8 boundaries before and 2 of 7 after the barrel. The GAA TFF is located within a large exon 2. An unduplicated GAA is preserved in all vertebrate species sequenced (GenBank accession numbers in parentheses): quail (BAA25890), mouse (NP_032090), and human (P10253) (see Fig. 6, which is published as supporting information on the PNAS web site). In contrast, the brush-border α-glucosidases seem always to be internally duplicated. This has been shown for SI from rabbit (A23945), rat (T10799), shrew (O62653), and human (P14410), and MGA from human (O43451; Figs. 3 and 6). The fourth α-glucosidase of the human is glucosidase II α (G2AN, alpha glucosidase II alpha subunit, GenBank accession no. NP_055425, chromosome 11), which serves in glucose trimming of Glc_1Man_o-glycosylated proteins within the endoplasmic reticulum and as an editor. G2AN shares only one exon boundary with MGA and SI (Figs. 3 and 5). These sequences and exon phases suggest that the duplicated mucosal luminal enzymes are an evolutionary adaptation for intestinal starch digestion by mammals.

Analysis of 105 different family 31 proteins in GenBank suggests that evolutionary divergence between the full N- and C-terminal brush-border enzyme domains occurred at a relative time of −8.3 but that specific MGAM and SI signatures I and II diverged at −4.6 and −6.2 before the present age. The oldest member of family 31 is *Caulobacter crescentus*, a proteobacterium that diverged at −11.

From the perspective of the TFF, the oldest member in GenBank is the *Xenopus laevis* egg envelope protein (gp37), which differentiated at −11.5. The MGA N terminus differentiated from the mammalian zonula protein (ZP) proteins at −9.5 before the present age. The family of mammalian TFF small peptides differentiated from the *Xenopus* TFF-like proteins at −5.7. These two relative evolutionary time scales suggest that MGA and SI duplication followed TFF development at −8.3 and specialization of these mammalian brush-border enzymes at −4 to −6 before the present age.

In the time scale of evolution, mammalian species are believed to have proliferated at the end of the Cretaceous period; amphibians developed in the Carboniferous period. These geological benchmarks suggest that the duplication and differenti-

ation of mammalian MGA and SI occurred $\approx 65 \times 10^6$ and of TFF $\approx 3 \times 10^8$ years ago, but that prokaryotic family 31 glucosidases originated at an earlier time (12).

Physiologic Deductions. The pattern of structural relatedness of the *MGAM* and *SI* genes is consistent with tandem duplication followed by divergent evolution with the structure of the gene evolving the most slowly, the protein sequence more quickly, and the chemical mechanism of the enzyme most rapidly (26). The complete conservation of a common proton donor and recipient site in all four terminal domains (in the pile-up in Fig. 5) suggests that differences in substrate binding must account for differing MGA and SI enzyme specificities. The functional collaboration of MGA and SI in terminal starch digestion, documented by the complementary enzyme specificities, is therefore anchored at the genomic level.

Craig Chinault, from the Baylor College of Medicine Human Genome Project, assisted with PAC clone screening. We also acknowledge the assistance of Patricia Aldred, who was funded by a Rank Prize Summer studentship. This work is a publication of the U.S. Department of Agriculture/Agricultural Research Service Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine and Texas Children's Hospital, Houston. This project was funded in part with federal funds from the U.S. Department of Agriculture/Agricultural Research Service (Cooperative Agreement 58-6250-1-003 to B.L.N.) and the Swiss National Foundation (Grant 3200-052736.97 to E.S.).

1. Myers, A. M., Morel, M. K., James, M. G. & Ball, S. G. (2000) *Plant Physiol.* **122,** 989–997.
2. Nichols, B. L., Eldering, J., Avery, S., Hahn, D., Quaroni, A. & Sterchi, E. E. (1998) *J. Biol. Chem.* **273,** 3076–3081.
3. Semenza, G., Auricchio, S. & Mantei, N. (2001) in *The Metabolic Basis of Inherited Disease*, eds. Scriver, C. R., Beaudet, A. L., Sly, W. S. & Valle, D. (McGraw–Hill, New York), 8th Ed., pp. 1623–1650.
4. Hassid, W. Z. & Ball, C. E. (1957) in *The Carbohydrates: Chemistry, Biochemistry, and Physiology*, ed. Pigman, W. (Academic, New York), pp. 501–506.
5. Whistler, R. L. & Corbet, W. M. (1957) in *The Carbohydrates: Chemistry, Biochemistry, and Physiology*, ed. Pigman, W. (Academic, New York), pp. 644–646.
6. Semenza, G., Auricchio, S. & Rubino, A. (1965) *Biochim. Biophys. Acta* **96,** 487–497.
7. Naim, H. Y., Sterchi, E. E. & Lentze, M. J. (1988) *J. Biol. Chem.* **263,** 19709–19717.
8. Genetics Computer Group (1994) *Program Manual for the Wisconsin Package* (Genetics Computer Group, Madison, WI), Version 9.
9. Molecular Biology Information Resource (1989) SAM: A Software Package for Sequence Assembly Management for UNIX Systems (Department of Cell Biology, Baylor College of Medicine, Houston), Version 8.
10. Altschul, S. F & Lipman, D. J. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 5509–5513.
11. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448.
12. Nei, M. & Kumar, S. (2000) *Molecular Evolution and Phylogenetics* (Oxford, New York), pp. 17–32, 300–301.
13. Corpet, F., Servant, F., Gouzy, J. & Kahn, D. (2000) *Nucleic Acids Res.* **28,** 267–269.
14. Chantret, I., Lacasa, M., Chevalier, G., Ruf, J., Islam, I., Mantei, N., Edwards, Y., Swallow, D. & Rousset, M. (1992) *Biochem. J.* **285,** 915–923.
15. Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. E. (2000) *J. Mol. Biol.* **299,** 501–522.
16. Higuchi, R. (1990) in *PCR Protocols: A Guide to Methods and Applications*, eds. Innis, M. A., Gelfand, D. H., Sninsky, J. J. & White, T. J. (Academic, San Diego), pp. 177–183.
17. Pischitzis, A., Hahn, D., Leuenberger, B. & Sterchi, E. E. (1999) *Eur. J. Biochem.* **261,** 421–429.
18. Nichols, B. L., Dudley, M. A., Nichols, V. N., Putman, M., Avery, S. E., Fraley, J. K., Quaroni, A., Shiner, M. & Carrazza, F. R. (1997) *Gastroenterology* **112,** 742–751.
19. Al Kazar, M., Desseaux, V. & Marchis-Mouren, G. (1998) *Eur. J. Biochem.* **252,** 100–107.
20. Frandsen, T. P. & Svensson, B. (1998) *Plant Mol. Biol.* **37,** 1–13.
21. Jesperson, H. M., MacGregor, E. A., Henrissat, B., Sierks, M. R. & Svensson, B. (1993) *J. Protein Chem.* **12,** 791–805.
22. Kashiwabara, S. I., Azuma, S., Tsuduki, M. & Suzuki, Y. (2000) *Biosci. Biotechnol. Biochem.* **64,** 1379–1393.
23. Okuyama, M., Okuno, A., Shimizu, N., Mori, H., Kimura, A. & Chiba, S. (2001) *Eur. J. Biochem.* **268,** 2270–2280.
24. Hoffmann, W., Jagla, W. & Wiede, A. (2001) *Histol. Histopathol.* **16,** 319–334.
25. de Souza, S. J., Long, M., Schoenbach, L., Roy, S. W. & Gilbert, W. (1997) *Gene* **205,** 141–144.
26. Janeček, S. (1996) *Protein Sci.* **5,** 1136–1143.