Sequence of the complete cDNA and the 5′ structure of the human sucrase-isomaltase gene

Possible homology with a yeast glucoamylase

Isabelle CHANTRET,*¶ Michel LACASA,†‡ Guillemette CHEVALIER,§ Jürg RUF,|| Ira ISLAM,* Ned MANTEI,|| Yvonne EDWARDS,* Dallas SWALLOW* and Monique ROUSSET§

*MRC Human Biochemical Genetics Unit, The Galton Laboratory, University College London, 4 Stephenson Way, London NW1 2HE, U.K., †IRSC Laboratoire sur les Virus et la Différenciation, 7 rue Guy Moquet, 94800 Villejuif, France, ‡Université Pierre et Marie Curie, 2 Place Jussieu, 75005 Paris, France, §Unité de Recherches sur la Différenciation Cellulaire Intestinale, INSERM U178, 16 avenue Paul-Vaillant Couturier, 94807 Villejuif Cedex, France, and ||Department of Biochemistry, Swiss Federal Institute of Technology, CH-8092 Zürich, Switzerland

The complete sequence of the 6 kb cDNA and the 5′ genomic structure are reported for the gene coding for the human intestinal brush border hydrolase sucrase-isomaltase. The human sucrase-isomaltase cDNA shows a high level of identity (83 %) with that of the rabbit enzyme, indicating that the protein shares the same structural domains in both species. In addition to the previously reported homology with lysosomal α-glucosidase, the sucrase and isomaltase subunits also appear to be homologous to a yeast glucoamylase. A 14 kb human genomic clone has been isolated which includes the first three exons and the first two introns of the gene, as well as 9.5 kb 5′ to the major start site of transcription. The first exon comprises 62 bp of untranslated sequence and the second starts exactly at the initiation ATG codon. Typical CAAT and TATA boxes are seen upstream of the first exon. A genetic polymorphism is described which involves a *Pst*I site in the second intron. Southern blotting, sequencing and mRNA studies indicate that the structures of the sucrase-isomaltase gene and its mRNA are unaltered in the two human colon cancer cell lines Caco-2 and HT-29 in comparison with normal human small intestine.

INTRODUCTION

The hydrolase sucrase-isomaltase (SI; EC 3.2.1.48 and 3.2.1.10) is expressed almost exclusively in the small intestinal brush border. It is a highly glycosylated protein which is synthesized as a single chain precursor (pro-SI) of apparent M_r 260000. After transport to the brush border membrane, it is cleaved proteolytically into two associated subunits, each of which has an active site with α-glucosidase activity of slightly different substrate specificity. The complex is anchored into the membrane via the *N*-terminal end of the isomaltase subunit (for reviews see [1,2]). The expression of SI is apparent in the human intestine at the 8th week of gestation, and after birth is maintained at significant levels only in the small intestine [3–5]. Within the small intestine, its expression varies along the crypt–villus axis, the enzyme being maximally expressed along the sides of the villi [6,7]. SI is also expressed in two human colon cancer cell lines HT-29 and Caco-2, but is repressed when these cell lines are cultured in conditions which lead to increased glucose utilization: Caco-2 cells treated with forskolin or monensin [8,9], and HT-29 cells cultured in 25 mM-glucose [10,11]. Using full-length rabbit [12] and partial human [13] SI cDNA clones as probes, a good correlation has been demonstrated between the expression of SI at the levels of mRNA and protein [5,14–17], suggesting regulation at the level of transcription. As an essential step towards characterizing the human SI gene and its regulatory elements, we have isolated further human cDNA clones to complete the RNA sequence and have characterized a 14 kb genomic clone. In addition to the homology between sucrase, isomaltase and the human lysosomal α-glucosidase already proposed [12,18,19], we demonstrate a likely homology with the recently cloned glucoamylase 1 from *Schwanniomyces occidentalis* [20]. We describe the 5′ structure of the human SI gene, including the first three exons and the 5′ flanking region.

MATERIALS AND METHODS

Libraries

Three different human cDNA libraries were used for the isolation of cDNA clones: a human jejunum λ gt11 library [13], a human intestinal λ ZAP library [21] and a λ gt11 library constructed with mRNA isolated from cultured Caco-2 cells on day 14. A human genomic library in EMBL3 vector, obtained from Dr. A. M. Frischauf (ICRF, London, U.K.) was used for the isolation of genomic clones.

Screening of libraries

The cDNA libraries were screened using the full-length rabbit SI cDNA [12] as probe. Positive clones which gave the strongest signal on plaque hybridization were isolated and characterized and designated I1–12 (from the λ gt11 jejunal library), C2 (from the λ gt11 Caco-2 library) and N4 (from the λ ZAP library). In order to complete the 5′ untranslated region, the jejunal λ gt11 library was rescreened using the 5′ *Apa*lI restriction fragment

Abbreviations used: SI, sucrase-isomaltase; DMEM, Dulbecco's modified Eagle's medium; SSC, 0.15 M-NaCl/0.015 M-sodium citrate; ORF, open reading frame; GRP, glucose-regulated protein; C/EBP, CAAT/enhancer binding protein.

¶To whom correspondence should be addressed, at: Unité de Recherches sur la Différenciation Cellulaire Intestinale, INSERM U178, Bât. INSERM, 16 avenue Paul-Vaillant Couturier, 94807 Villejuif Cedex, France.

The nucleotide sequence data reported will appear in the EMBL, GenBank and DDBJ Nucleotide Sequence Databases under accession no. X63597.

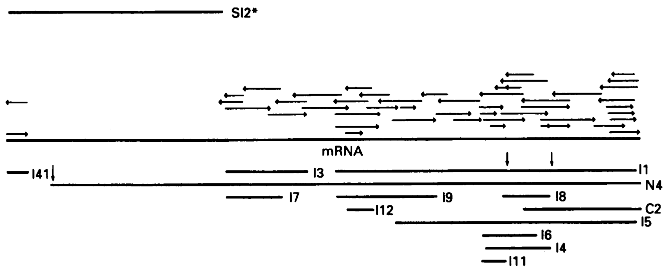

Fig. 1. Cloning and sequencing strategy for human SI cDNA clones isolated from three independent human cDNA libraries

The sizes and positions of the 13 characterized clones are shown in relation to the full-length mRNA (middle line). The I clones were isolated from the jejunal λ gt11 library, the N4 clone was isolated from the intestinal λ ZAP library, and the C2 clone was from the Caco-2 λ gt11 library. Arrows indicate the direction and the length of the sequences. The previously published SI 2 clone is indicated by an asterisk [13]. Vertical arrows indicate insertion positions of anomalous stretches of sequence.

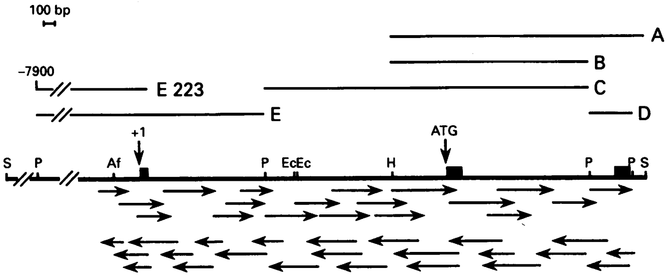

Fig. 2. Sequencing strategy and partial restriction map for the 14 kb human SI genomic clone

Five restriction fragments [A (2 kb), B (1.8 kb), C (2.6 kb), D (0.4 kb) and E (9 kb)] are indicated. The subclone E223, obtained after Exo nuclease III/mung bean nuclease treatment of the fragment E, was used for the 5′ terminus mapping with RNAase. The positions of the transcription start site (+1) and of the ATG codon are indicated. Arrows indicate the direction and the length of the sequences. The middle line shows a partial restriction map of the genomic clone 4.11.e. and the 5′ structure of the human SI gene, including the size and position of the three exons: Af, *Afl*II; Ec, *Eco*RI; H, *Hin*dIII; P, *Pst*I; S, *Sal*I.

(268bp) of SI2 [13] as probe. The six clones isolated were characterized by double digestions with *Eco*RI/*Hae*III and *Eco*RI/*Apa*II and hybridization to an oligonucleotide complementary to nucleotides 84–101 in the cDNA sequence. The clone I41 was selected for further analysis because it appeared to be the longest in the 5′ direction.

The human EMBL3 genomic library was screened with the SI2 cDNA probe. All the cDNA probes were ^{32}P-labelled according to Feinberg & Vogelstein [22] using a multiprime DNA labelling kit (Amersham International), and oligonucleotides were end-labelled using [γ-^{32}P]ATP as described by Suggs *et al.* [23].

Sequencing and analysis

Inserts from the cDNA clones were subcloned into pUC13, Bluescript, M13 or pTZ18R vectors. The strategy for sequencing and characterization of the cDNA clones is outlined in Fig. 1. The clones isolated from the λ gt11 library range in size from 250 bp (I41) to 3200 bp (I1) and encompass most of the cDNA apart from nt 2874–3137. This gap was filled by sequencing the clone N4 isolated from the λ ZAP library. Analysis of this clone also confirmed the sequence across the *Eco*RI site which resides at nt 2093/4 (the 3′ end of SI2 and the 5′ end of I3 and I7). The λ gt11 clone I41 contains 50 more 5′ untranslated nucleotides than does SI2. A few anomalous stretches of sequence were identified in some clones but not others. These are indicated in Fig. 1, and probably reflect incomplete splicing as described for example in cystic fibrosis transmembrane conductance regulator cDNA by others [24].

Fig. 2 shows the sequencing strategy for the genomic clone 4.11.e, including a partial restriction map. Five insert restriction fragments (A, B, C, D and E) were subcloned in both orientations into pTZ18R. Nested deletions of the subclones C and E of the genomic clone were obtained by the exonuclease III/mung bean nuclease procedure [25].

Sequencing was carried out by the dideoxy chain termination method [26] using [^{35}S]dATP [27] and Klenow DNA polymerase or Sequenase on double- or single-stranded DNA. Much of the cDNA and small sections of the genomic clones were sequenced using oligonucleotide primers. The oligonucleotides were synthesized on an Applied Biosystems 391 DNA synthesizer PCR-mate. All the cDNA and genomic sequences have been determined on both strands.

Computer analysis of sequences

The sequence analyses were carried out using the Seqnet facility at the Daresbury Laboratory, Daresbury, Warrington, U.K. with the University of Wisconsin GCG programs, the BISANCE programs [28,29] and the PC GENE package from IntelliGenetics for the search of homologies in the EMBL Databank (release 25) and the SWISS-PROT Databank (release 17).

Tissue specimens and cultured cells

Normal human ileum was resected from an irreversibly brain-damaged organ donor cadaver. Human colon adenocarcinoma Caco-2 cells were cultured in Dulbecco's modified Eagle's minimum essential medium (DMEM) (Eurobio, les Ulis, France) with 20 % heat-inactivated fetal calf serum (Boehringer–Mannheim) and 1 % non-essential amino acids (Gibco). Control cells or cells treated for 48 h with 25 μM-forskolin (France-Biochem) or 1 μM-monensin (Calbiochem Behring, La Jolla, CA, U.S.A.) were used after 14 days in culture [14,15,30]. Human colon adenocarcinoma HT-29 cells [9,11] were cultured with (Glc+) or without (Glc−) 25 mM-glucose in DMEM (Eurobio) with 10 % heat-inactivated and dialysed fetal calf serum. The cells were used after 18 days in culture. The Glc− cells were used after eight passages in the glucose-free medium.

RNA analysis

Total RNA was extracted from normal human ileum or from Caco-2 or HT-29 cells using the guanidium isothiocyanate/CsCl method [31]. Poly(A)+ RNAs were prepared using oligo(dT)-cellulose chromatography.

For blot analysis, 10 μg of each total or poly(A)+ RNA specimen was run on a 1 % agarose gel after denaturation in 1 M-glyoxal [32], transferred to a Hybond N (Amersham) membrane and hybridized with the ^{32}P-labelled cDNA clone I1. The filter was washed twice in 2 × SSC/0.1 % SDS at room temperature, once in 0.1 × SSC/0.1 % SDS at 50°C and then in 0.1 × SSC/0.1 % SDS at 70°C for 15 min (where SSC is standard sodium citrate).

Mapping of the 5′-terminus of SI mRNA was performed by two different techniques. Primer extension of Caco-2 and ileum mRNA was performed with an oligonucleotide complementary to nucleotides 101–84 of the SI cDNA sequence. The extended DNA was electrophoresed through a 6 % polyacrylamide/urea sequencing gel alongside a sequence of pTZ18R, primed with the M13 universal primer, as a length marker (^{35}S labelling). The

```
                                     tattttggcagccttatccaagtctggtacaacatagcaaagagaacaggctatgaaataag     62
ATGGCAAGAAAGAAATTTAGTGGATTGGAAATCTCTCTGATTGTCCTTTTTGTCATAGTTACTATAATAGCTATTGCCTTAATTGTTGTTTTAGCAACTA    162
AGACACCTGCTGTTGATGAAATTAGTGATTCTACTTCAACTCCAGCTACTACTCGTGTGACTACAAATCCTTCTGATTCAGGAAAATGTCCAAATGTGTT    262
AAATGATCCTGTCAATGTGAGAATAAACTGCATTCCAGAACAATTCCCAACAGAGGGAATTTGTGCACAGAGAGGCTGCTGCTGGAGGCCGTGGAATGAC    362
TCTCTTATTCCTTGGTGCTTCTTCGTTGATAATCATGGTTATAACGTTCAAGACATGACAACAACAAGTATTGGAGTTGAAGCCAAATTAAACAGGATAC    462
CTTCACCTACACTATTTGGAAATGACATCAACAGTGTTCTCTTCACAACTCAAAATCAGACACCCAATCGTTTCCGGTTCAAGATTACTGATCCAAATAA    562
TAGAAGATATGAAGTTCCTCATCAGTATGTAAAAGAGTTTACTGGACCCACAGTTTCTGATACGTTGTATGATGTGAAGGTTGCCCAAAACCCATTTAGC    662
ATCCAAGTTATTAGGAAAAGCAACGGTAAAACTTTGTTTGACACCAGCATTGGTCCCTTAGTGTACTCTGACCAGTACTTACAGATCTCAGCCCGTCTTC    762
CAAGTGATTATATTTATGGTATTGGAGAACAAGTTCATAAGAGATTTCGTCATGATTTATCCTGGAAAACATGGCCAATTTTTACTCGAGACCAACTTCC    862
TGGTGATAATAATAATAATTTATACGGCCATCAAACATTCTTTATGTGTATTGAAGATACATCTGGAAAGTCATTCGGTGTTTTTTTAATGAATAGCAAT    962
GCAATGGAGATTTTTATCCAGCCTACTCCAATAGTAACATATAGAGTTACCGGTGGCATTCTGGATTTTTACATCCTTCTAGGAGATACACCAGAACAAG   1062
TAGTTCAACAGTATCAACAGCTTGTTGGACTACCAGCAATGCCAGCATATTGGAATCTTGGATTCCAACTAAGTCGCTGGAATTATAAGTCACTAGATGT   1162
AGTGAAAGAAGTGGTAAGGAGAAACCGGGAAGCTGGCATACCATTTGATACACAGGTCACTGATATTGACTACATGGAAGACAAGAAAGACTTTACTTAT   1262
GATCAAGTTGCGTTTAACGGACTCCCTCAATTTGTGCAAGATTTGCATGACCATGGACAGAAATATGTCATCATCTTGGACCCTGCAATTTCCATAGGTC   1362
GACGTGCCAATGGAACAACATATGCAACCTATGAGAGGGGAAACACACAACATGTGTGGATAAATGAGTCAGATGGAAGTACACCAATTATTGGAGAGGT   1462
ATGGCCAGGATTAACAGTATACCCTGATTTCACTAATCCAAACTGCATTGATTGGTGGGCAAATGAATGCAGTATTTTCCATCAAGAAGTGCAATATGAT   1562
GGACTTTGGATTGACATGAATGAAGTTTCCAGCTTTATTCAAGGTTCAACAAAAGGATGTAATGTAAACAAATTGAATTATCCACCGTTTACTCCTGATA   1662
TTCTTGACAAACTCATGTATTCCAAAACAATTTGCATGGATGCTGTGCAGAACTGGGGTAAACAGTATGATGTTCATAGCCTCTATGGATACAGCATGGC   1762
TATAGCCACAGAGCAAGCTGTACAAAAAGTTTTTCCTAATAAGAGAAGCTTCATTCTTACCCGCTCAACATTTGCTGGATCTGGAAGACATGCTGCTCAT   1862
TGGTTAGGAGACAATACTGCTTCATGGGAACAAATGGAATGGTCTATAACTGGAATGCTGGAGTTCAGTTTGTTTGGAATACCTTTGGTTGGAGCAGACA   1962
TCTGTGGATTTGTGGCTGAAACCACAGAAGAACTTTGCAGAAGATGGATGCAACTTGGGGCATTTTATCCATTTTCCAGAAACCATAATTCTGACGGATA   2062
TGAACATCAGGATCCTGCATTTTTTGGGCAGAATTCACTTTTGGTTAAATCATCAAGGCAGTATTTAACTATTCGCTACACCTTATTACCCTTCCTCTAC   2162
ACTCTGTTTTATAAAGCCCATGTGTTTGGAGAAACAGTAGCAAGACCAGTTCTTCATGAGTTTTATGAGGATACGAACAGCTGGATTGAGGACACTGAGT   2262
TTTTGTGGGGCCCTGCATTACTTATTACTCCTGTTCTAAAACAGGGAGCAGATACTGTGAGTGCCTACATCCCTGATGCTATTTGGTATGATTATGAATC   2362
TGGTGCAAAAAGGCCATGGAGGAAACAACGGGTTGATATGTATCTTCCAGCAGACAAAATAGGATTACATCTTAGAGGAGGTTATATCATCCCCATTCAA   2462
GAACCAGATGTAACAACAACAGCAAGCCGTAAGAATCCTCTAGGACTTATAGTCGCATTAGGTGAAAACAACACAGCCAAAGGAGACTTTTTCTGGGATG   2562
ATGGAGAAACTAAAGATACAATACAAAATGGCAACTACATATTATATACATTTTCAGTTTCTAATAACACATTAGATATTGTGTGCACACATTCATCATA   2662
TCAGGAAGGAACTACCTTAGCATTTCAGACTGTAAAAATCCTTGGGTTGACAGACAGTGTTACAGAAGTTAGAGTGGCGGAAAATAATCAACCAATGAAC   2762
GCTCATTCCAATTTCACTTATGATGCTTCTAACCAGGTTCTCCTAATTGCAGATCTCAAACTTAATCTTGGAAGAAACTTTAGTGTTCAATGGAATCAAA   2862
TTTTCTCAGAAAATGAAAGATTTAATTGTTATCCAGATGCAGATTTGGCAACTGAACAAAAGTGCACACAACGTGGCTGTGTATGGAGAACGGGTTCTTC   2962
TCTATCCAAAGCACCTGAGTGTTACTTTCCCAGACAAGATAACTCTTATTCAGTCAACTCAGCTCGCTATTCATCCATGGGTATAACAGCTGACCTCCAA   3062
CTAAATACTGCAAATGCCAGAATAAAGTTACCTTCTGACCCCATCTCAACTCTTCGTGTGGAGGTGAAATATCACAAAAATGATATGTTGCAGTTTAAGA   3162
TTTATGATCCCCAAAAGAAGAGATATGAAGTACCAGTACCGTTAAACATTCCAACCACCCCAATAAGTACTTATGAAGACAGACTTTATGATGTGGAAAT   3262
CAAGGAAAATCCTTTTGGCATCCAGATTCGACGGAGAAGCAGTGGAAGAGTCATTTGGGATTCTTGGCTGCCTGGATTTGCTTTTAATGACCAGTTCATT   3362
CAAATATCGACTCGCCTGCCATCAGAATATATATATGGTTTTGGGGAAGTGGAACATACAGCATTTAAGCGAGATCTGAACTGGAATACTTGGGGAATGT   3462
TCACAAGAGACCAACCCCCTGGTTACAAACTTAATTCCTATGGATTTCATCCCTATTACATGGCTCTGGAAGAGGAGGGCAATGCTCATGGTGTTTTCTT   3562
ACTCAACAGCAATGCAATGGATGTTACATTCCAGCCAACTCCTGCTCTAACTTACCGTACAGTTGGAGGGATCTTGGATTTTTATATGTTTTTGGGCCCA   3662
ACTCCACAAGTTGCAACAAAGCAATACCATGAAGTAATTGGCCATCCAGTCATGCCAGCTTATTGGGCTTTGGGATTCCAATTATGTCGTTATGGATATG   3762
CAAATACTTCAGAGGTTCGGGAATTATATGACGCTATGGTGGCTGCTAACATCCCCTATGATGTTCAGTACACAGACATTGACTACATGGAAAGGCAGCT   3862
AGACTTTACAATTGGTGAAGCATTCCAGGACCTTCCTCAGTTTGTTGACAAAATAAGAGGAGAAGGAATGAGATACATTATTATCCTGGATCCAGCAATT   3962
TCAGGAAATGAAACAAAGACTTACCCTGCATTTGAAAGAGGACAGCAGAATGATGTCTTTGTCAAATGGCCAAACACCAATGACATTTGTTGGGCAAAGG   4062
TTTGGCCAGATTTGCCCAACATAACAATAGATAAAACTCTAACGGAAGATGAAGCTGTTAATGCTTCCAGAGCTCATGTAGCTTTCCCAGATTTCTTCAG   4162
GACTTCCACAGCAGAGTGGTGGGCCAGAGAAATTGTGGACTTTTACAATGAAAAGATGAAGTTTGATGGTTTGTGGATTGATATGAATGAGCCATCAAGT   4262
TTTGTAAATGGAACAACTACTAATCAATGCAGAAATGACGAACTAAATTATCCACCTTATTTCCCAGAACTCACAAAAAGAACTGATGGATTACATTTCA   4362
GAACAATTTGCATGGAAGCTGAGCAGATTCTTAGTGATGGAACATCAGTTTTGCATTACGATGTTCACAATCTCTATGGATGGTCACAGATGAAACCTAC   4462
TCATGATGCATTGCAAAAGACAACTGGAAAAAGAGGGATTGTAATTTCTCGTTCCACGTATCCTACTAGTGGACGATGGGGAGGACACTGGCTTGGAGAC   4562
AACTATGCACGATGGGACAACATGGACAAATCAATCATTGGTATGATGGAATTTAGTCTGTTTGGAATATCATATACTGGAGCAGACATCTGTGGTTTTT   4662
TCAACAACTCAGAATATCATCTCTGTACCCGCTGGATGCAACTTGGAGCATTTTATCCATACTCAAGGAATCACAACATTGCAAATACTAGAAGACAAGA   4762
TCCCGCTTCCTGGAATGAAACTTTTGCTGAAATGTCAAGGAATATTCTAAATATTAGATACACCTTATTGCCCTATTTTTACACACAAATGCATGAAATT   4862
CATGCTAATGGTGGCACTGTTATCCGACCCCTTTTGCATGAGTTCTTTGATGAAAAACCAACCTGGGATATATTCAAGCAGTTCTTATGGGGTCCAGCAT   4962
TTATGGTTACCCCAGTACTGGAACCTTATGTTCAAACTGTAAATGCCTACGTCCCCAATGCTCGGTGGTTTGACTACCATACAGGCAAAGATATTGGCGT   5062
CAGAGGACAATTTCAAACATTTAATGCTTCTTATGACACAATAAACCTACATGTCCGTGGTGGTCACATCCTACCATGTCAAGAGCCAGCTCAAAACACA   5162
TTTTACAGTCGACAAAAACACATGAAGCTCATTGTTGCTGCAGATGATAATCAGATGGCACAGGGTTCTCTGTTTTGGGATGATGGAGAGAGTATAGACA   5262
CCTATGAAAGAGACCTATATTTATCTGTACAATTTAATTTAAACCAGACCACCTTAACAAGCACTATATTGAAGAGAGGTTACATAAATAAAAGTGAAAC   5362
GAGGCTTGGATCCCTTCATGTATGGGGGAAAGGAACTACTCCTGTCAATGCAGTTACTCTAACGTATAACGGAAATAAAAATTCGCTTCCTTTTAATGAA   5462
GACACTACCAACATGATATTACGTATTGATCTGACCACACACAATGTTACTCTAGAAGAACCAATAGAAATCAACTGGTCATGAagatcaccatcaattt   5562
tagttgtcaatgggaaaaaacaccaggatttaagtttcacagcacttacaattttccctcttcacttggttcttgtactctacaaaatatagctttcata   5662
acatcgaaaagttattttgtagcgtacatcaatgataatgctaattttattatagtaatgtgacttggattcaattttaaggcatatttaacaaaatttg   5762
aatagccctatttatccttgttaagtatcagctacaattgtaaactagttactaaacatgtatgtaaatagctaagatataatttaaacgtgattttaa   5862
attaaataaaattttatgtaattatatatactatattttctcaatgtttagcagatttaagatatgtaacaacaattatttgaagatttaattacttc   5962
ttagtatgtgcatttaattagaaaaagagaataaaaaatgtaagtgtaaaaaaaaaaa  6021
```

Fig. 3. cDNA sequence of the human SI gene

The first in-frame start (ATG) and stop (TGA) codons, as well as the polyadenylation signals, are underlined.

DNA was then transfered on to the Gene screen as described by Cebrian *et al.* [33]. After u.v. immobilization, the filter was hybridized with the ^{32}P-labelled 5′ *Apa*I fragment from SI2. For mapping with RNAase [34] the subclone E223 (−7900 to +69) in pTZ18R vector was linearized with *Afl*II (see Fig. 2). A ^{32}P-labelled antisense RNA corresponding to nucleotides −230 to +69 was transcribed with T7 polymerase. Ileum or Caco-2 poly(A)+ RNA (5 μg) was hybridized with the labelled antisense

```
H I  MARKK----FSGLEISLIV------LFVIVTIIAIALIVV-LATKTPAVDEISDS----TSTPATTRVTTNPSDSGKCPNVLND  69
H S  -----------------------NQIFSE-------------------------------------------------------  6
Y G  M---------------IFLKL------IKSIVIGLGLVSAIQAAPASSIG------------SSASASSSSESSQATIPNDVTL  51

H I  PVNVRINCIPE-QFPTEGICAQRGCCWRPWNDSLI------PWCFFVDNH-GYNVQDMTTTSIGVEAKLNRIPSPT---LFGND  142
H S  --NERFNCYPDADLATEQKCTQRGCVWRTGSSLSKAPE-----CYFPRQDNSYSVNSARYSSMGITADLQLNTANARIKLPSDP  83
Y G  GVKQIPNIFNDSAVDANA------------------------------AAKGYDLVNVTNTPRGLTGILKLKEA---TNIYGYD  102
           *   -    -                                    * -     -  *-   *    -     -

H I  INSVLFTTQNQTPNRFRFKITDPN-NRRYEVPHQY-VKEFTGPTVSDTLYD-----VKVAQNPFSIQVIRKSNGKTLFDTSIGP  219
H S  ISTLRVEVKYHKNDMLQFKIYDPQ-KKRYEVPVPLNIPTTPISTYEDRLYD-----VEIKENPFGIQIRRRSSGRVIWDSWLPG  161
Y G  FDYLNLTVEYQADTRLNVHIEPTDLSDVFVLPEHLVVKPLVEGDAQSYNFDNSDLVEEYSNTDFSFEVIRSSTKEVLFSTKGNP  186
        -               *        - -*    -            -*            *   - * *    -- -

H I  LVYSDQYLQISARLPSDYIY-GIGEQVHKRFRHDLSWKTWPIFTRDQLPGDNNNNLYGHQTFFMCIEDTSGKSFGVFLMNSNAM  302
H S  FAFNDQFIQISTRLPSEYIY-GFGEVEHTAFKRDLNWNTWGMFTRDQPPGYKLNS-YGFHPYYMALEEE-GNAHGVFLLNSNAM  242
Y G  LVFSNQFIQFNSSLPKNHVITGLGESIHGLVNEPGSVKT--LFAND-VGDPIDGNIYGVHPVYLDQRYDTETTHAVYWRTSAIQ  267
       -  *--*  - **   -  * **  *          *  -*- *          **    --        -  *-   *

H I  EIFIQPTPIVTYRVTGGILDFYILLGDTPEQVVQQYQQLVGLPAMPAYWNLGFQLSRWNYKSLDVVKEVVRRNREAGIPFDTQV  386
H S  DVTFQPTPALTYRTVGGILDFYMFLGPTPQVATKQYHEVIGHPVMPAYWALGFQLCRYGYANTSEVRELYDAMVAANIPYDVQY  326
Y G  EVLIGEES-ITWRALSGVIDLYFFSGPTPKDAIQQYVKEIGLPAFQPYWSLGYHQCRWGYDTIEKLSEVVENFKKFNIPLETIW  350
     --       -*-*   *--* *   * **     **   -* *    ** **-   *- *     - *-        ** -  -

H I  TDIDYMEDKKDFTYDQVAFNGLPQFVQ--DLHDHGQKYVIILDPAI--SIGRRANGTTYATYERGNTQHVWINESDGSTPIIGE  466
H S  TDIDYMERQLDFTIGE-AFQDLPQFVD--KIRGEGMRYIIILDPAI--SGN---ETKTYPAFERGQQNDVFVKWPNTNDICWAK  402
Y G  SDIDYMDSYKDFTYDPHRFPLDEYRKFLDELHKNNQHYVPILDAAIYVPNPNNATDNEYQPFHYGNETDVFLKNPDGSLYI-GA  433
     -*****-   ***     *           -      *- *** **            *  -  *-   *--

H I  VWPGL---------------------TVYPDFTNPNCIDWWANECSIFHQE-VQYDGLWIDMNEVSSFIQGST-KGCNVNKLN  526
H S  VWPDLPNITIDKTLTEDEAVNASRAHVAFPDFFRTSTAEWWAREIVDFYNEKMKFDGLWIDMNEPSSFVNGTTTNQCRNDELN  485
Y G  VW------------------------QVTLFSRFLSRKHSDMDKVIKDWYELTPFDGIWADMNEVSSFCVGSCGTGKYFENPA  492
     **                                                *   -**-* **** ***  *-

H I  YPPFTPDILDKL--MYSKTICMDAVQNW--------------------------------------------------------  552
H S  YPPYFPELTKRTDGLHFRTICMEAEQIL--------------------------------------------------------  513
Y G  YPPFTVGSKATSYPVGFDVSNASEWKSIQSSISATAKTSSTSSVSSSSSTIDYMNTLAPGKGNINYPPAYIYNMQGDSDLATHA  576
     ***-          -

H I  ---------GK---QYDVHSLYGYSMAIATEQAVQKVFPNKRSFILTRSTFAGSGRHAAHWLGDNTASWEQMEWSITGMLEFSL  624
H S  -------SDGTSVLHYDVHNLYGWSQMKPTHDALQKT-TGKRGIVISRSTYPTSGRWGGHWLGDNYARWDNMDKSIIGMMEFSL  589
Y G  VSPNATHADGTVE--YDIHNLYGYLQENATYHALLEVFPNKRPFMISRSTFPRAGKWTGHWGGDNTADWAYAYFSIPQAFSMGI  658
              *     **-* ***-     *  *-      **  ---***-  -*-   ** *** * *     **       -

H I  FGIPLVGADICGFVAETTEELCRRWMQLGAFYPFSRNFNSDGYEHQDPAFFGQNSLLVKSSRQYLTIRYTLLPFLYTLFYKAHV  708
H S  FGISYTGADICGFFNNSEYHLCTRWMQLGAFYPYSRNHNIANTRRQDPASWNET--FAEMSRNILNIRYTLLPYFYTQMHEIHA  671
Y G  AGLPFFGADVCGFNGNSDSELCSRWMQLGSFFPFYRNHNYLGAIDQEPYVWES---VAEATRTSMAIRYLLLPYYYTLLHESHT  739
      *-   ***-***   -   ** ******-*-*- ** *      *-*  -         -*  - *** ***- **     *

H I  FGETVARPVLHEFYEDTNSWIEDTEFLWGPALLITPVLKQGADTVSAYIPDAIWYDYES--------------GAKRP---WRK  775
H S  NGGTVIRPLLHEFFDEKPTWDIFKQFLWGPAFMVTPVLEPYVQTVNAYVPNARWFDYHT---------------GKDIGV--RG  738
Y G  TGLPILRAFSWQFPNDRSLSGVDNQFFVGDGLVVTPVLEPGVDKVKGVFPGAGKEEVYYDWYTQ------------REVHFKDG  811
      *  - *     *  -         *  *   --****      *    * *   -

H I  QRVDMYLPADKIGLHLRGGYIIPIQEPDVTTTASRKNPLGLIVALGENNTAKGDFFWDDGETKDTIQNGNYILYTFSVSNNTLD  859
H S  QFQTFNASYDTINLHVRGGHILPCQEPAQNTFYSRQKHMKLIVAADDNQMAQGSLFWDDGESIDTYERDLYLSVQFNLNQTTLT  822
Y G  KNETLDAPLGHIPLHIRGGNVLPTQEPGYTVAESRQNPFGLIVALDNDGKAQGSLYLDDGESLVV---DSSLLVSFSVSDNTLS  892
                * **-*** --* ***      **     ****      * *  - ****-         -   * -   **

H I  IVCTHSSYQEGTTLAFQTVKILGLTDSVTEVRVAENNQPMNAHSNFTYDASNQVLLIA--DL--KLNLGRNFSVQW-  931
H S  STILKRGYINKSETRLGSLHVWGKGTTPVNAVTLTYNGNKNSLP-FNEDTTNMILRIDLTTH--NVTLEEPIEINWS  896
Y G  AS-PSGDY--KADQPLANVTILGVGHKPKSVKF------ENANVDFTYKKST-VFVTGLDKYTKDGAFSKDFTITW-  958
            *   -      - - *                 *-   *    -  -                   - *
```

Fig. 4. Comparison of amino acid sequences of human isomaltase (H I), human sucrase (H S), and yeast glucoamylase 1 (Y G)

Amino acids identical in the three sequences are indicated by ★; conservative amino acids in the three sequences are indicated by –. The WXDMNE (Trp-Xaa-Asp-Met-Asn-Glu) motif is in bold.

RNA. Single-stranded material was hydrolysed with RNAases A and T1, and after denaturation the protected labelled RNA was run on a 6% polyacrylamide/urea sequencing gel along with the pTZ18R sequence primed with a 22-mer reverse primer, as size markers.

Southern blot analysis

Genomic DNA isolated from normal human intestine, HT-29 and Caco-2 cells was *Eco*RI-digested and analysed by Southern blotting using standard procedures [34]. The filter was probed with a mix of ^{32}P-labelled cDNA clones corresponding to the full-length cDNA (Fig. 1). For the analysis of the genetic polymorphism, genomic DNA samples were digested with *Pst*I and subjected to electrophoresis on 0.8% agarose for 22 h. The DNA was transferred passively on to Hybond N. The filters were probed with the genomic subclones C and (in some cases) D (Fig. 2) labelled with ^{32}P by random primer labelling, and washed down to 2 × SSC/0.1% SDS at 65°C for 30 min.

RESULTS

Isolation of cDNA clones

In order to isolate the complete cDNA for human SI and to obtain a longer sequence in the 5′ untranslated region, we have characterized 13 further cDNA clones from three independent libraries (cf. Fig. 1). The resulting combined nucleotide sequence of these clones is presented in Fig. 3. The sequence around the ATG codon (the A is designated as nt 63, Fig. 3) includes the

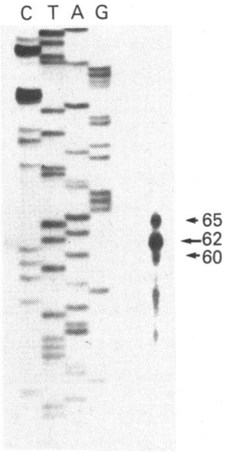

Fig. 5. Mapping of the 5′ terminus of human SI mRNA with RNAase

Poly (A)$^+$ RNA (5 μg) from normal human ileum was hybridized with labelled antisense RNA transcribed from the E223/ pTZ18R plasmid digested with *Afl*II (see Fig. 2). A sequencing reaction on pTZ18R was loaded as size marker. The numbers indicate the sizes (nt) of the protected RNA fragments obtained after RNAase treatment.

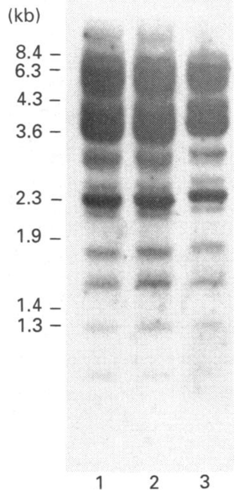

Fig. 6. Southern blot analysis of *Eco*RI-digested DNA (20 μg) from normal human tissue (lane 1), Caco-2 cells (lane 2) and HT-29 (Glc+) cells (lane 3)

The probe used corresponds to the full-length human SI cDNA.

most highly conserved nucleotides, A and G at positions 60 and 66 respectively, of the consensus sequence described by Kozak [35]. The open-reading frame (ORF) of human SI, of which 2030 nt have been published previously [13] is now complete : it is 5481 nt long and encodes a 1827-amino-acid peptide precursor. Downstream of the stop codon (TGA), two potential polyadenylation signals (AATAAA) are identified at positions 5867 and 5992.

Comparative analysis of the human and rabbit [12] cDNA sequences reveals 83 % identity overall: 85 % in the 26 bp of 5′ untranslated sequence available for comparison, 74 % in the 3′ untranslated region and 84 % in the coding region. Within the coding region, 206 single base pair substitutions are in the first position of the codon, 156 in the second position and 502 in the third position. As in the rabbit, the sequence comprises two parts, one coding for the isomaltase subunit (nt 63–2855) and one coding for the sucrase subunit (nt 2856–5543), which show 37.7% identity; an additional 34% of the sequence shows conservative changes in the deduced amino acid sequence. Both subunits contain the active site sequence Trp-Ile-Asp-Met-Asn-Glu, as in the rabbit. Eighteen potential *N*-glycosylation sites are present compared with 19 in the rabbit, and 12 of these are conserved between the two species (three in the isomaltase subunit and nine in the sucrase subunit).

We have compared the peptide sequences of human and rabbit sucrase and isomaltase with that of human lysosomal α-glucosidase [19] and various other enzymes which hydrolyse similar substrates: these included a bacterial isoamylase [36] and two yeast enzymes, invertase [37] and glucoamylase 1 [20]. The only new apparent homology was observed with the peptide sequence of glucoamylase 1. The Clustal program [38] was used to align all six sequences (sucrase and isomaltase, from both human and rabbit, human lysosomal α-glucosidase and yeast glucoamylase 1). Parameter settings which optimized the alignments of five of the sequences, were not optimal for that of the glucoamylase 1. Glucoamylase 1 was therefore realigned by eye, starting from positions 468 to 473, which appeared to be equivalent to the Trp-Ile-Asp-Met-Asn-Glu motif in SI. Apart from *N*-terminal domain and three gaps, this alignment showed two blocks of highly conserved sequences, at positions 189–430 and 500–836 in the human isomaltase subunit. In the region where all six sequences can be aligned (from amino acid 72 to the end of the isomaltase peptide sequence), 137 amino acids are identical in all six enzymes and a further 128 identical in five out of six. In these cases, a conservative amino acid substitution is usually found in the sixth sequence. Since this analysis is cumbersome to present, we show in Fig. 4 just the comparison of the human sucrase and isomaltase subunits with yeast glucoamylase 1.

Identification of the transcription start site

The 5′ end of SI mRNA was mapped by primer extension of mRNA extracted from normal human ileum or from Caco-2 cells harvested on day 14. A faint band of 101 bases was observed which represented an initiation site 62 bp upstream of the ATG codon (result not shown). RNAase protection assays were also used to identify the transcription start site. mRNA extracted from ileum or from Caco-2 cells was hybridized with an antisense ^{32}P-labelled RNA probe corresponding to nucleotides -230 to $+69$ in the genomic sequence, which includes 237 bases upstream from the 5′ end of the I41 cDNA clone and seven bases of the first intron. The major band protected from RNAase digestion was 62 nt in length, and minor bands of 65 and 60 nt were also seen (Fig. 5).

5′ structure of the gene and analysis of the 5′ flanking region

Hybridization of the SI probes of *Eco*RI digested Caco-2, HT-29 or normal human DNA revealed various bands ranging in size from 0.7 kb to 13 kb (Fig. 6). This allowed us to make an estimate on the size of the SI gene, which must be at least 55 kb.

The genomic clone described here contains an insert 14 kb in length and covers the 5′ end of the gene. The sequence of the first three exons and the first two introns is shown in Fig. 7. The first 62 bp exon, which corresponds to the 5′ untranslated region of the cDNA, is followed by a 2424 bp intron. The second 118 bp exon begins exactly at the ATG codon and is followed by a 1224 bp intron and the third 137 bp exon. The junctions between the introns and exons are in agreement with the published consensus sequences [39].

Two potentially interesting sequences were detected within the first intron. One between nucleotides 1953 and 1961 could represent a binding site for CAAT/enhancer binding protein (C/EBP) [40]; the other between nucleotides 2010 and 2023 is

```
aatagaaagaactacttttatgttcttaagatattaatagttcacagctttgagaaatcaaagagtatctgacagtacaattactaattaacttagattc - 155
ccagagagaaacatttatgtaaactactttcagggttaaggcttttaacaattgtatgttgagcagaagattattaaacactgataggctggtgagggtg -  55
caataaaactttatgagtaggtcaatatatacctaagcagctttccttgatcatTATTTTGGCAGCCTTATCCAAGTCTGGTACAACATAGCAAAGAGAA     46
CAGGCTATGAAATAAGgtaagtctatggctcatatgtttttgttattttatgtttcaactaattatatttttcttaatttgtgttagcagtattcc    146
catctatctgcactaggctaaaatagcaatatcatatatttaagttgccatttgaaatatgtagtaccttgtattaaacaggaaaaaatatttctaatt    246
tttttttaaatacaggatttttttggcccaactcttcattataactttgtaagacttaggtggaatcatagcaagttttgtttaatgactctatatga    346
aagaaaatgagtaccgattaagatttgaaaaatagatgaataataaagtatgcttaaatatatttaagaaaatctatataaaatagaatgctacattt    446
ataaaacattactgttttctagctgaccttttaaagatcattcatcaagttgttatacaataaagtattattcatcttcataataaatatgcagagaaag    546
ggtatcaatatatattttccaagaagagttgcttaaaattactgtaatccaaaattacagggatcatattttttatcacagtaattttaagcaattgtt    646
ttaaattttgttttgattttcctagacaaaggagagtttgtttcatgacaaaaatgcatttttaagacttcccttatcaaaaaggaaagcacaattctc    746
attattgacttcttggtagcaaagtaggttatcacttcaaatgagatttacacttatttaaagttgtacactaaatgtaactttgtcatttctcaacata    846
acatcacgcaactgggcttcatttttcgaacagctttgtatgttacctgttgctctgtttactttacattcatttttctttgaaaattactgatgcaataa    946
acagcttacatatctggaacagaaaattgtatactaagaataaatgctaatctgcagatattcatttataagctatactaaaaaattaagttttgctta   1046
cttaatattattctctgactactttataattttgctaacaagaatgtaacgatctaattgaggaaattaacatatacataaataattgcaatggacat   1146
ttaatatgctatgctaaatgaagaaaggctagattaactgtgcctcagagtggtaaggaactcaccatggcatttgaattcagcaaaatttagaattcac   1246
ctagaagagtggatggggagggattgaaattgaacagcaggtatgaaggcaatatgttatctaaaaattatgctgtggtgctggggataaaatgaaggca   1346
taattctaaaaatactttgaagataaagtcagatggacttattgaatgattgaatgttgtaaaacatatggaagaaggaaaaaaaaggtgtttaggttgg   1446
gacaactgggtacatgagttgacatagttacagaacagggaccattgtgagaccaaaaaaaaaaaaaaaacaaaccgtgatttaaacccaataattta   1546
ttaaagcacttactgtagtaattctaagttctactttggtatatgtctttgaattaagatttatactaaaatcatagaatacagataatccatggaaact   1646
tgagaactaagaaaaataagatatatgtggtcaaaattaaattaaacaaatatgctttcagataattattggagatgtgttctttcctgtctacatttgg   1746
gtatcttccatatgctacctcaaaagtataatgtcttttaataataaatatgtatttttcgatataagagttttgtatatctcatttgctgctgtttgat   1846
gaaggaacttggtgaaaattcaaaagatttcgaaacatagctagcaactcccaaatatatactacgctatcaaatatatgtatttatgttccttttcga   1946
caatattgtggaaagtagtgttggaattagagagaggagagttcaagaaaaaaactgaagttttaactttgaactataacttattctttgttaagctta   2046
aagttcctgatcagtaaaatgtgaataactatttccttcatagaattgttgtgagaattgagaaacagtgcaagtggaagcctttgccttgcaaaatgct   2146
ctaaatgagaaggcaagacacacctacttgaagatcttttttccatctgtattttcatcttagtgtactactactactgtaatgtgcatcagtaattca   2246
ttcctttactttaaattgcatgtgttgcttacatttatatcagtataagtttcaggttattaagtacatgcttaatatattattaagtaagtgcttaa   2346
tttaggatattatgtaagtgctgtcttcaactctgcattatacatagtaaaaaaaagaatatgtaatttcatgagaattgttggtggtttattatagtta   2446
tatattttacatttaaatatatattcttttctgttttagATGGCAAGAAAGAAATTTAGTGGATTGGAAATCTCTCTGATTGTCCTTTTTGTCATAGT   2546
TACTATAATAGCTATTGCCTTAATTGTTGTTTTAGCAACTAAGACACCTGCTGTTGATGgtaagtagtacattaaaagcatggctacattatcgtgaagt   2646
taggaggaagttactccacaatagttataggttttaaaaaataatctgtgtaagatgaaacacttatagaaaaatccagggtaagttacacaattacttc   2746
aactttttctattcattatattccatatttcaatgatgagctattgtaattaaggaatccatgtgctttgacattaattgccttccctagctacttgaca   2846
aagagcaaagaggcaaagggatcaaatcgttaatgtaaggtggaataatcgcctactggtgtgagcatcatctcattagtgggaaaacctgggttttgc   2946
tccagccttcacacacttgcaaaggaaataaggatttagcaaatattctttgttctggttagattttattttaaggagagagagaggttgtatgagaaa   3046
ggcaggtagaggaaagttgaaaaaagacttggatagtttaattatctttttctgaatattttaattctaatttagagttggtacatacggaatgctgaat   3146
ctctttcttcattattctctcttgagccaatcagtattgtaaaggatgttacacattatcttctagactacagttagcatttcctattcctgttaatctt   3246
ttgttttacctacttccagaaactgagagcttactgtattatgaagcaatatattttaatgatatctcattgctggaatgttatttatatggcgttggta   3346
tctaccattcttcgctattccagctttctctactttacctatcaaattattcccttgcaagggataatttgaataactttcacctactttcctctctct   3446
aatttcttctcaccttctctatatgtttttcataactcatattgctccatcttttcttcagctataaaatacaagttgtaaatacagtatgaatgtatat   3546
tttgttaccttctgaatctacaccaagtcctttgtaaatatattttgtatttctgcagaaacagtcaataggattctacactgtcaattttaaaagtat   3646
gattggatctcttccttttgatgactagtgcttgaaattgtgggtctttttaaatgcaaaatttattttaaaattttcattttatatgaattcatgtatt   3746
ctttcaaaaagtacttattacttgataattctgagaaaattaatttaatgtcatgtttattttattgagttttcctcccccagAAATTAGTGATTCTAC   3846
TTCAACTCCAGCTACTACTCGTGTGACTACAAATCCTTCTGATTCAGGAAAATGTCCAAATGTGTTAAATGATCCTGTCAATGTGAGAATAAACTGCATT   3946
CCAGAACAATTCCCAACAGAGgtctgcag  3975
```

Fig. 7. The 5′ structure of the human SI gene

The 5′ upstream region and the introns are typed in lower case letters. The CAAT and TATA boxes are underlined. A potential binding site for C/EBP and the region of similarity with intestinal fatty acid binding protein are indicated with broken lines. The vertical arrow indicates an ATG codon; horizontal arrows indicate the limits of the antisense RNA used for the mapping of the 5′ terminus.

very similar (12 nucleotides out of 14) to a consensus sequence found in the 5′ non-transcribed region of two genes expressed in intestine, the intestinal fatty acid binding protein [41] and the cellular retinol binding protein II [42] genes.

In the 5′ upstream region, of which 6.3 kb have been sequenced, typical CAAT and TATA boxes were seen at positions −106 and −29 respectively. Since SI and human lysosomal α-glucosidase are apparently derived from the same ancestral gene [19,43] the sequences of both promoter regions were compared, but no similarity was found.

Genetic polymorphism within intron 2

Southern blots of DNA samples from six or more random unrelated individuals digested with *Eco*RI, *Taq*I, *Msp*I, *Hae*III, *Hin*fI or *Pst*I were probed with the genomic subclones C or D to search for commonly genetically determined polymorphism (Fig.

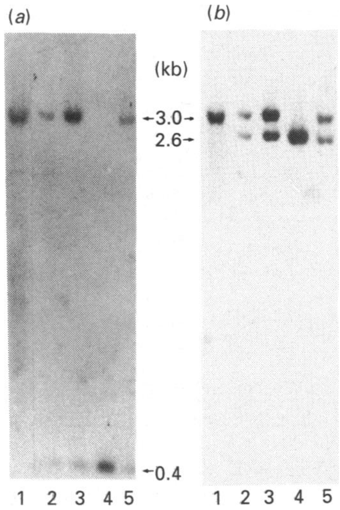

Fig. 8. *Pst*I polymorphism detected with genomic probes SI D (*a*) and SI C (*b*)

*Pst*I digested DNA from five individuals were transferred on to Hybond N membrane and probed with SI C, followed by reprobing with SI D. The SI phenotypes of the individuals are: 1, SI C1; 2, SI C2-1; 3, SI C2-1; 4, SI C2; 5, SI C2-1.

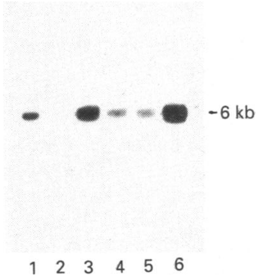

Fig. 9. Northern blot analysis

Poly (A)$^+$ (lanes 1 and 2) or total (lanes 3–6) RNA was extracted from Glc− (lane 1) or Glc+ (lane 2) HT-29 cells, normal human ileum (lane 3), control Caco-2 cells (lane 6) or Caco-2 cells treated for 48 h, with monensin (lane 4) or forskolin (lane 5).

8). A polymorphism involving a *Pst*I site was detected with both probes, but the other enzymes failed to detect common variation. Three common *Pst*I phenotypes were detected, which could be attributed to allelic variation at the *Pst*I site which separates subclone C from D (nt 1116). In the rarer type I allele (frequency 0.36) this site is absent, resulting in a 3.0 kb fragment which is detectable by both clones C and D. In the more common allele (frequency 0.64) the site is present (as in our genomic clone), giving a 2.6 kb band detectable with subclone C and a 0.4 kb band detectable with subclone D. Parents of 34 families from the CEPH (Centre d'Etude du Polymorphisme Humain) series were typed and 16 families, comprising 207 individuals, were tested for linkage purposes. SI was shown to be closely linked to a number of anonymous DNA markers on chromosome 3 [43a].

Northern blot analysis of SI mRNA

As described previously [13–15], Northern blot analysis of RNA extracted from normal human ileum and from differentiated Caco-2 cells shows high levels of a single hybridizing species of 6 kb (Fig. 9). The same 6 kb mRNA was detected in preparations from Caco-2 cells treated with forskolin or monensin (lanes 4 and 5), but at much lower levels than in the Caco-2 control cells. The 6 kb transcript was also detected at a relatively low level in the differentiated HT-29 cells cultured in the absence of glucose (Fig. 9, lane 1). When HT-29 cells were cultured in the presence of glucose, a condition in which no SI activity can be found, no SI mRNA was detected (Fig. 9, lane 2).

DISCUSSION

The cDNA sequence of human SI is described in this paper: it includes the entire coding region (5481 nucleotides), and the complete 5′ and 3′ untranslated regions. Of the two potential polyadenylation signals, the first is not followed by a GT-rich sequence which has been reported to be a second signal recognized by the eukaryotic machinery for the maturation of the 3′ end [44,45]. It seems likely that this first AATAAA is used infrequently or not at all, since the polyadenylation starts at position 5947 in all the 3′ clones that we have analysed.

The deduced amino acid sequences of the two subunits of human SI have been compared with those of rabbit SI and the high level of similarity indicates that the protein structure of human SI is likely to be the same as that proposed for the rabbit enzyme [12]. The peptide sequence of human SI precursor contains 18 putative *N*-glycosylation sites, 12 of which are conserved between human and rabbit. The knowledge of the entire peptide sequence and of the *N*-glycosylation sites will be particularly useful for the study of SI deficiency, where the absence of expression of this enzyme is often associated with a blockage in its transport and with glycosylation abnormalities (for review see [1]).

On the basis of the similarity of their amino acid sequences, it has previously been hypothesized that the SI and human lysosomal α-glucosidase genes have evolved from a common ancestor [12,18,19]. Here we report apparent homology with another α-glucosidase, glucoamylase 1 from yeast. The active site motif Trp-Ile-Asp-Met-Asn-Glu, as well as the other stretches of amino acid sequence which are conserved between the six peptides, are absent in all the other entries of the SWISS-PROT Databank (release 17) which includes about 35 sequences of enzymes with related catalytic function. This indicates that these motifs could constitute markers for the genes derived from the same ancestor.

By Southern blot studies, we have estimated the size of the total human SI gene to be at least 55 kb. The SI gene in HT-29 and Caco-2 cells shows the same pattern of *Eco*RI restriction fragments as normal human genomic DNA, indicating no major rearrangement within or around the gene in these tumour cells.

We have isolated and characterized a 14 kb genomic segment of the 5′ end of the human SI gene which contains the first three exons, the two first introns and the beginning of the third intron. A genetic polymorphism was detected which involves a *Pst*I site in the second intron. Analysis of the families agreed with the previous assignment of the gene to chromosome 3.

The major transcription start site determined by two different techniques (primer extension and RNAase mapping) was found to be located only 7 bp upstream of the longest cDNA clone in the 5′ untranslated region (I41). Typical consensus sequences, TATA and CAAT, were seen −29 and −106 nt respectively upstream of this major transcription start site. The first exon is only 62 bp long, and is followed by a 2.5 kb intron which surprisingly ends exactly at the ATG codon. The presence of this large intron could possibly explain the weakness of the signals obtained in the primer extension experiments, since we have evidence that many of the SI RNA molecules selected by oligo(dT) from ileum or Caco-2 cells are incompletely spliced (I. Chantret, G. Chevalier & M. Rousset, unpublished work).

No sequence similarity in the 5′ flanking regions could be found between the acid α-glucosidase and SI genes. Indeed, the nucleotide composition of this region is quite different: the α-glucosidase promoter regions is very GC-rich [43], which is a characteristic of a house-keeping gene, while that of SI is AT-rich. Unlike the acid α-glucosidase, which is an ubiquitous lysosomal enzyme, SI is restricted to the enterocytes of the small intestine. Very little is known about the sequences involved in the regulation of expression of intestinal proteins, although experiments with transgenic mice suggest that sequences allowing correct cellular expression of intestinal fatty acid binding protein reside within a 277 bp sequence 5′ to the first exon [46]. Within this region of the intestinal fatty acid binding protein gene, and also repeated upstream in the 5′ flanking region, resides a conserved sequence (TGAACTTTGAACTT) [41] which has also been identified in other genes expressed in the intestine (retinol binding protein and apolipoprotein A) [41,42]. It is therefore of some interest that this sequence is found within the first intron of the human SI gene. In this intron there is also a putative binding site for C/EBP, a protein which appears to have a role in terminal differentiation in a number of tissues [47].

In addition to searching for regulatory sequences involved in tissue specificity, further studies will be required to elucidate the mechanisms responsible for the catabolic glucose repression of SI in Caco-2 and HT-29 cells (Fig. 9). The significance of such regulation *in vivo* is suggested by results which show decreased SI expression in diabetic rats [48]. A number of other genes are also repressed by glucose, including yeast invertase [49] (which also hydrolyses sucrose), yeast glucoamylase [20] (which seems to derive from the same ancestral gene as SI (see Fig. 2), yeast alcohol dehydrogenase II [50] and the human glucose-regulated proteins (GRP) [51]. In the cases of the yeast invertase and alcohol dehydrogenase and the human GRP proteins, glucose-sensitive 5′ regulatory elements and consensus binding sequences have been identified [52–54]. We were unable to find either of these consensus binding sequences in the 5′ region of the human SI gene. However, it is noteworthy that the GRP consensus sequence was also not found in the (glucose transporter 1) gene GLUT1, which is co-repressed with GRP by glucose [55].

In most of the studies described in this paper, the two human colon carcinoma cell lines Caco-2 and HT-29 were examined in parallel with the normal human small intestine, since chromosomal rearrangements are frequent in colorectal carcinomas [56,57] and tumour cell lines. All the evidence suggests that the SI gene is unaltered in these cancer cells and that the SI RNA is processed in the same way as in the human small intestine. This lends support to the suitability of these cells as a model for future studies designed to identify the regulatory elements of the human SI gene.

We thank Wendy Pratt and David Matthews for their continuous help and support and Jacques Armier, Clare Harvey and Philip Johnson for preparing the oligonucleotides. We are also grateful to Thierry Foulon for help with computer analysis of sequencing data. This work was financed in part by grant from ARC (6127), GEFLUC, the Fondation pour la Recherche Medicale and NATO (0789/88). I.C. was the recipient of fellowships from INSERM and Fondation pour la Recherche Médicale.

REFERENCES

1. Hauri, H. P. (1988) in Subcellular Biochemistry (Harris, J. R., ed.) vol. 12, pp. 155–219, Plenum Press, New York
2. Semenza, G. (1989) Biochem. Int. **18**, 15–33
3. Zweibaum, A., Triadou, N., Kedinger, M., Augeron, C., Robine-Leon, S., Pinto, M., Rousset, M. & Haffen, K. (1983) Int. J. Cancer **32**, 407–412
4. Lacroix, B., Kedinger, M., Simon-Assmann, P., Rousset, M., Zweibaum, A. & Haffen, K. (1984) Early Human Dev. **9**, 95–103
5. Sebastio, G., Hunziker, W., O'Neill, B., Malo, C., Menard, D., Auricchio, S. & Semenza, G. (1987) Biochem. Biophys. Res. Commun. **149**, 830–839
6. Nordstrom, C. & Dahlquist, A. (1973) Scand. J. Gastroenterol. **8**, 407–416
7. Skovbjerg, H. (1981) Biochem. J. **193**, 887–890
8. Rousset, M., Laburthe, M., Pinto, M., Chevalier, G., Rouyer-Fessard, C., Dussaulx, E., Trugnan, G., Boige, N., Brun, J.-L. & Zweibaum, A. (1985) J. Cell. Physiol. **123**, 377–385
9. Rousset, M., Trugnan, G., Brun, J.-L. & Zweibaum, A. (1986) FEBS Lett. **208**, 34–38
10. Pinto, M., Appay, M. D., Simon-Assmann, P., Chevalier, G., Dracopoli, N., Fogh, J. & Zweibaum, A. (1982) Biol. Cell **44**, 193–196
11. Zweibaum, A., Pinto, M., Chevalier, G., Dussaulx, E., Triadou, N., Lacroix, B., Haffen, K., Brun, J.-L. & Rousset, M. (1985) J. Cell. Physiol. **122**, 21–29
12. Hunziker, W., Spiess, M., Semenza, G. & Lodish, H. F. (1986) Cell. **46**, 227–234
13. Green, F., Edwards, Y., Hauri, H. P., Povey, S., Ho, M. W., Pinto, M. & Swallow, D. (1987) Gene, **57**, 101–110
14. Chantret, I., Trugnan, G., Dussaulx, E., Zweibaum, A. & Rousset, M. (1988) FEBS Lett. **235**, 125–128
15. Rousset, M., Chantret, I., Darmoul, D., Trugnan, G., Sapin, C., Green, F., Swallow, D. & Zweibaum, A. (1989) J. Cell. Physiol. **141**, 627–635
16. Chandrasena, G., Sunitha, I., Nanthakumar, N. & Henning, S. J. (1990) J. Cell. Biol. **111**, 218a (Abstr.)
17. Traber, P. G. (1990) Biochem. Biophys. Res. Commun. **173**, 765–773
18. Hoefsloot, L. H., Hoogeveen-Westerveld, M., Kroos, M. A., van Beeumen, J., Reuser, A. J. J. & Oostra, B. A. (1988) EMBO J. **7**, 1697–1704
19. Martiniuk, F., Mehler, M., Tzall, S., Meredith, G. & Hirschhorn, R. (1990) DNA Cell Biol. **9**, 85–94
20. Dohmen, R. J., Strasser, A. W. M., Dahlems, U. M. & Hollenberg, C. P. (1990) Gene. **95**, 111–121
21. Mantei, N., Villa, M., Enzler, T., Wacker, H., Boll, W., James, P., Hunziker, W. & Semenza, G. (1988) EMBO J. **7**, 2705–2713
22. Feinberg, A. P. & Vogelstein, B. (1983) Anal. Biochem. **132**, 6–13
23. Suggs, S. V., Wallace, R. B., Hirose, T., Kawashima, E. H. & Itakura, K. (1981) Proc. Natl. Acad. Sci. U.S.A. **78**, 6613–6617
24. Riordan, J. R., Rommens, J. M., Kerem, B. S., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J.-L., Drumm, M. L., Annuzzi, M. C., Collins, F. S. & Tsui, L.-C. (1989) Science **245**, 1066–1073
25. Henikoff, S. (1984) Gene. **28**, 351–359
26. Sanger, F., Nicklen, S. & Coulsen, A. R. (1977) Proc. Natl. Acad. Sci. U.S.A. **74**, 5463–5467
27. Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) Proc. Natl. Acad. Sci. U.S.A. **80**, 3963–3965
28. Gouy, M., Gautier, C., Attimonelli, M., Lanave, C. & Di Paola, G. (1985) Compr. Appl. Biosciences. **1**, 167–172
29. Dessen, P., Fondrat, C., Valencien, C. & Mugnier, C. (1990) Compr. Appl. Biosci. **6**, 355–356
30. Pinto, M., Robine-Leon, S., Appay, M. D., Kedinger, M., Triadou, N., Dussaulx, E., Lacroix, B., Simon-Assmann, P., Haffen, K., Fogh, J. & Zweibaum, A. (1983) Biol. Cell. **47**, 323–330
31. Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. & Rutter, W. J. (1979) Biochemistry **18**, 5294–5299
32. Thomas, P. S. (1980) Proc. Natl. Acad. Sci. U.S.A. **77**, 5201–5205
33. Cebrian, J., Berthelot, N. & Laithier, M. (1989) J. Virol. **63**, 523–531
34. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
35. Kozak, M. (1987) Nucleic. Acids Res. **15**, 8125–8148
36. Chen, J. H., Chen, Z. Y., Chow, T. Y., Chen, J. C., Tan, S. T. & Hsu, W. H. (1990) Biochim. Biophys. Acta **1087**, 309–315
37. Perlman, D. & Halvorson, H. O. (1981) Cell **25**, 525–536
38. Higgins, D. G. & Sharp, P. M. (1988) Gene **73**, 237–244
39. Mount, S. M. (1982) Nucleic. Acids Res. **10**, 459–472
40. Graves, B. J., Johnson, P. F. & McKnight, S. L. (1986) Cell **44**, 565–576
41. Sweetser, D. A., Birkenmeier, E. H., Klisak, I. J., Zollman, S., Sparkes, R. S., Mohandas, T., Lusis, A. J. & Gordon, J. I. (1987) J. Biol. Chem. **262**, 16060–16071

42. Demmer, L. A., Birkenmeier, E. H., Sweetser, D. A., Levin, M. S., Zollman, S., Sparkes, R. S., Mohandas, S. T., Lusis, A. J. & Gordon, J. I. (1987) J. Biol. Chem. **262**, 2458–2467

43. Hoefsloot, L. H., Hoogeveen-Westerveld, M., Reuser, A. J. J. & Oostra, B. A. (1990) Biochem. J. **272**, 493–497

43a. Swallow, D., Islam, I., Attwood, J., Harvey, C., Chantret, I., Lacasa, M., Chevalier, G. & Rousset, M. (1991) Cytogenet. Cell Genet. **58**, 1881 (abstr.)

44. Taya, Y., Devos, R., Tavernier, J., Cheroutre, H., Engler, G. & Fiers, W. (1982) EMBO J. **1**, 953–958

45. McLauchlan, J., Gaffney, D., Whitton, J. L. & Clements, J. B. (1985) Nucleic Acids Res. **13**, 1347–1368

46. Sweetser, D. A., Hauft, S. M., Hoppe, P. C., Birkenmeier, E. H. & Gordon, J. I. (1988) Proc. Natl. Acad. Sci. U.S.A. **85**, 9611–9615

47. Umek, R. M., Friedman, A. D. & McKnight, S. L. (1991) Science **251**, 288–292

48. Najjar, S. M., Hampp, L. T., Rabkin, R. & Gray, G. M. (1991) Am. J. Physiol. **260**, G275–G283

Received 30 August 1991/3 February 1992; accepted 14 February 1992

49. Carlson, M. (1987) J. Bacteriol. **169**, 4873–4877

50. Denis, C. L., Ciriacy, M. & Young, E. T. (1981) J. Mol. Biol. **148**, 355–368

51. Shiu, R. P. C., Pouyssegur, J. & Pastan, I. (1977) Proc. Natl. Acad. Sci. U.S.A. **74**, 3840–3844

52. Chang, S. C., Erwin, A. E. & Lee, A. S. (1989) Mol. Cell. Biol. **9**, 2153–2162

53. Nehlin, J. O. & Ronne, H. (1990) EMBO J. **9**, 2891–2898

54. Thukral, S. K., Eisen, A. & Young, E. T. (1991) Mol. Cell. Biol. **11**, 1566–1577

55. Wertheimer, E., Sasson, S., Cerasi, E. & Ben-Neriah, Y. (1991) Proc. Natl. Acad. Sci. U.S.A. **88**, 2525–2529

56. Vogelstein, B., Fearon, E. R., Kern, S. E., Hamilton, S. R., Preisinger, A. C., Nakamura, Y. & White, R. (1989) Science **244**, 207–211

57. Muleris, M., Salmon, R. I. & Dutrillaux, B. (1990) Cancer Genet. Cytogenet. **46**, 143–156