Comparison of two yeast invertase genes: conservation of the upstream regulatory region

Laura Sarokin and Marian Carlson*

Department of Human Genetics and Development, and Institute for Cancer Research, Columbia University, College of Physicians and Surgeons, New York, NY 10032, USA

Received 14 June 1985; Revised and Accepted 8 August 1985

ABSTRACT

The yeast genome contains a dispersed family of invertase structural genes (SUC1-SUC5, SUC7). Five of these genes are located very close to telomeres and are flanked by large regions of homologous sequence; recombination between telomeres could account for the dispersal of these SUC genes to different chromosomes. The SUC2 locus, in contrast, is not near a telomere and does not share large regions of flanking homology with the other loci. We examine here the relationship between SUC2 and one of the telomeric genes, SUC7. Sequence comparison revealed homology extending from about position -624 to +1791, which is close to the end of the mRNA. The 5' noncoding sequence includes two highly conserved regions: the region between -140 and +1, which contains the TATA box and presumably other promoter elements, and a second region extending from -508 to -400, which corresponds to the upstream regulatory region.

INTRODUCTION

The genome of the yeast Saccharomyces contains a dispersed family of structural genes for invertase, the SUC (sucrose) genes (SUC1 through SUC5 and SUC7). All of the SUC genes are homologous within the coding region (1, 2). Analysis of the physical structures of different SUC⁺ loci has shown, moreover, that all of the SUC genes except for SUC2 share large regions of homologous flanking sequences and are located very close to telomeres (3). These SUC genes and their flanking sequences are located towards the telomere from the telomere-adjacent X sequences characterized by Chan and Tye (4). Most yeast strains do not carry SUC⁺ genes at all of the telomeric SUC loci; rather, they have zero, one or a few SUC⁺ genes. Cloning of the natural negative alleles, called suc^o alleles, has shown that suc^o loci correspond to telomeres lacking the SUC⁺ gene and flanking sequences (3). The evidence suggests that dispersal of the

telomeric SUC genes to different chromosomes occurred by rearrangements of chromosome termini, probably as a result of homologous recombination between repeated sequences present at different telomeres (3).

The SUC2 gene is an unusual member of the family with respect to its physical structure. Blot hybridization analysis indicated that the homology between the SUC2 locus and other SUC loci is limited to the structural gene and close flanking sequences (3). Moreover, although SUC2 appears genetically to reside near the end of chromosome IX (5), analysis of cloned DNA from the SUC2 locus has shown that the SUC2 gene is located at least 14 kb from telomere-adjacent sequences (3). The SUC2 locus also differs from the telomeric SUC loci in that most, if not all, strains carry SUC gene information at the SUC2 locus, either a SUC2⁺ gene or a mutant gene, or pseudogene (1, 2, 6). The evolutionary relationship between SUC2 and the other SUC genes is thus very different from the relationship among the other, telomeric copies.

The SUC2 gene has previously been characterized in detail. It encodes two forms of invertase: a secreted, highly glycosylated enzyme and an intracellular, nonglycosylated form. The two enzymes are synthesized from two differently regulated mRNAs with different 5' ends; the 1.9-kb mRNA encoding secreted invertase includes a signal sequence and its level is regulated by glucose repression (7, 8). The other, 1.8-kb mRNA does not include the entire signal sequence and is produced constitutively at a low level (7, 8). We have previously identified a regulatory region between 418 and 650 nucleotides upstream from the coding sequence that controls the expression of secreted invertase in response to glucose availability (9). This region has been shown to confer regulated expression to a heterologous gene and to mediate the regulatory effects of trans-acting positive and negative regulators of SUC2 (10).

We here examine the relationship between SUC2 and one of the telomeric SUC genes, SUC7. All of the telomeric SUC genes are homologous throughout the relevant region, and SUC7 was arbitrarily chosen as a representative of this group. We determined the nucleotide sequence of the regions flanking SUC7,

and also the 5' and 3' ends of the coding sequence, for comparison to the SUC2 sequence. We have previously determined the sequence of the SUC2 gene and flanking regions (11).

Comparison of the sequences of SUC2 and SUC7 showed that the coding region and the 5' and 3' noncoding sequences close to the coding region are highly conserved. In addition, the regulatory region far upstream from the SUC2 gene has an homologous counterpart upstream from SUC7.

MATERIALS AND METHODS

Construction of plasmids

The SUC7 gene and flanking DNA were cloned previously (3). The following subclones were constructed by standard procedures (12). Plasmid pRT62-30 was constructed by inserting the 0.9-kb Eco RI-Bam HI fragment spanning the 5' end of the coding sequence into pBR322. Plasmid pRT85-3 carries the 3 kb Hind III fragment containing the 5' noncoding region inserted into the Hind III site of pBR322. pRT131-10 was constructed by subcloning the 1.5-kb Xba I fragment spanning the 3' end of the gene into the Xba I site of pLB50 (L. Babiss and H. Ginsberg, Columbia University, unpublished). The Xba I fragment was sometimes prepared directly from DNA of a phage carrying the 8-kb Eco RI fragment on which the SUC7 gene is located (3).

Nucleotide sequence analysis

Sequence analysis was carried out on 5' end-labeled restriction fragments by the procedure of Maxam and Gilbert (13) as previously described (11). In all cases, fragments labeled at a unique end were obtained by cleaving with a second restriction enzyme and separating the resulting fragments. Additional sequence of the SUC2 region, not previously reported, between positions -784 and -637 was determined by sequencing from a Xho I site, located at the point of a deletion removing sequences between -1900 and -870 (9), towards the SUC2 gene.

Sequence alignment

Regions of homology within the 5' noncoding sequences of SUC2 and SUC7 were identified with a dot matrix program provided by S. Goff. The best alignment of the two sequences was determined with the program ALIGN (14).

Yeast strains

The genotypes of yeast strains used in this study are as follows: DBY615 (MAT α SUC7 ura3), obtained from F. Lacroute via D. Botstein; MCY130 (MAT α SUC7 ade2 gal2); MCY257 (MAT α SUC7 ade2); MCY501 (MAT α SUC2 ade2 gal2); MCY1207 (MAT α SUC7 suc2- Δ 9 ura3). MCY501 is isogenic to S288C (MAT α SUC2 gal2). MCY130, MCY257 and MCY1207 are congenic to S288C as a result of at least ten crosses to strains of the S288C genetic background.

Invertase assays

Glucose-repressed and derepressed cultures were prepared as described by Celenza and Carlson (15). To obtain glucose-repressed cells, cultures were grown to exponential phase in rich medium containing 2% glucose. Derepressed cells were prepared by shifting repressed cells to rich medium containing 0.05% glucose for 2.5 h. Secreted invertase was assayed (16) in whole cells as described (15). The gel assay was carried out as described before (17).

RESULTS

Nucleotide sequence analysis

Figure 1 shows a restriction map of the SUC7 gene indicating the restriction sites and strategy used for sequence analysis. Nucleotide sequence analysis was performed on restriction fragments prepared from clones described in Methods. The sequences of the 5' and 3' regions of SUC2 and SUC7 are compared in Figures 2 and 3, respectively. For purposes of clarity in the ensuing discussion, all nucleotide positions will refer to positions in the SUC2 sequence. Thus, for SUC7, the numbers are those of the corresponding nucleotide in the SUC2 sequence; the actual positions in the SUC7 sequence can be ascertained by reference to Figures 2 and 3.

Coding region

The two genes are highly conserved within the coding region. Like SUC2, the SUC7 gene encodes both secreted and intracellular invertase (6). The translation initiation codons for the precursor to secreted invertase (position +1) and for the intracellular invertase (position +61) are conserved between SUC2 and SUC7 (Figure 2). Within the first 222 nucleotides of the

Figure 1. Strategy for sequencing the SUC7 region. The map of the SUC7 gene shows restriction sites used for sequence analysis; other sites are not shown. Arrows indicate the direction and extent of the nucleotide sequence read from each 5' end-labeled site. The open bar represents the coding region.

amino-terminal coding region, there are only 24 differences, which result in only six amino acid changes. One of these changes occurs at position +19 within the signal sequence; SUC2 encodes leucine and SUC7, asparagine. At the 3' ends of the coding regions, there is complete conservation of the final 69 nucleotides through the termination codon. Secreted invertase is a glycoprotein, and three tripeptide sequences that may serve as recognition sites for addition of asparagine-linked polysaccharide chains (18, 19) were conserved in both proteins: the asparagine-X-threonine sequences at positions +67, +190 and +1534.

5' noncoding region

The homology between SUC2 and SUC7 in the 5' noncoding region extends upstream from the structural gene to position -220; further upstream is a gap in the homology due to 38 nucleotides present in SUC2 but not SUC7. The overall homology within this region, excluding the four small gaps, is nearly 80%; however, the homology is 90% between -1 and -140 and only about 60% between -141 and -220, although this latter region contains small patches of very good homology. Previous deletion analysis has shown that nucleotides -140 through -418 are not required for high-level derepression of secreted invertase (9); thus, essential promoter elements must lie downstream of -140, within the region that is highly conserved. Interestingly, the TATA box at -133 (8, 9) is not perfectly conserved; the SUC2 sequence is TATAAATA, and the SUC7 sequence is TACAAATA. The nucleotide at -40 corresponding to the 5' end of the 1.9-kb SUC2 mRNA encoding

```

SUC2  GGAGGTTTCCCAATGAACAAAGGACAGGGGACGGTGAGCTGTGGAAG
      *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
SUC7  TGCACAGTGAAACTACTTCTAGATGCGTAATTCCTGGAAGGACAAACG
      *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
                                           -700

      -700
GTATCCATTTTATCATGTTTCGTTTGTACAAGCAGCAGCATCTAAGACATTACCGTATGGGAGTTGTTCTCTAGCTAGTCTCTCGCTCCCCAGCAA
*  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
GAAGAGGAAAAAAGAGGACACACGATAGTAAAAGACATCTAAAATTTCTTTGATTCAAACGTGATCTCTCACTTTGATGTCCACTGAATTCCTTT
*  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
                                           -600

      -600
GCTCAAAAAGTACGTCATTAGAATAGTTTGTGAGCAAAATTACCGTCCGTATGCTACGTTAGAA---AGGCCACAGTATTCTTACCAAAGGCGTG
*  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
TCCGATGTCAAACATCATTATAGCAA-TCAGTGAAGCAATCTGGAGAGAGTGATTCAATTTATAATTTAGGAATCCCGCATGTTGCAAGAAAGACAGTA
*  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
                                           -500

      -500
CCT--TTGTTGAAGTCGATCCATTATGAGGGCTTCCATTATT---CCCGCATTTTATTACTCTGAACAGGAATAAAAAGAAAAAACCCAGTTTAGGAA
*  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
GGTGCTGAATATATCTTACTTAAAGATGAAATAGTCCGTTTTTTTCTCCGCTTTTTTCTCAGCTGAGCGGAAATAAAAATTTGTC-----TTCGAAA
*  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
                                           -400

      -400
ATTATCCGGGGGCGAAGAAATACCGGTAGCGTTAATCGACCCAC----GTCCAGGTTTTTCCATGAGGTTTCTGGAAAACTGAGGAGGAATGTGA
*****  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
ATTATCCGGGG-CAAGAAATGGCGTAGGATTTCTCAACCCCACTTTTTGTGAAGTTTTTCCCGGAAATCTGTGGAAACAATATTGTTGCGCAATGA
*  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
                                           -300

      -300
TTATAATCCCTTATGTGATGTCTAAGACITTTAAGGTACGCCGATGTTGCTTATTACCATCATAGAGAGCTTTCTTTTCGAGGAATGCTTAAACGA
****  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
TTATTGAGCCTATTTCACA-----GATTATTCAGGTACAGATACCTTCTACGTAAAAAAAAGTGATATTGCCTGTCAAGAAAA-----
*  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
                                           -200

      -200
CTTTGTTGACAAAAATGTTGCCTAAGGGCTCTATAGTAACCATTTGGAAGAAAGATTGACGACITTTT---TTTTTGGATTTCGATCCTATAATCC
*  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
-----CTCAATACTAACAATCTGGAAGAGGGTGAGAGAAATAAATGACTTTTTTGGGCTGTGATCATAGAATTC
*  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
                                           -100

      -100
--TTCCTCTGAAAGAAACATATAAATAGATATGTATTATTCTTC---AAACATTCTCTTGTCTTGTGCTTTTTTTTACCATATATCTTACTTTT
*  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
CATTTTCTCTCAAGAGAAACGTACAATAGATATGTATTGTTCTTTTAAACATTCTCTTGTCTTGTGCTTTTTTTTCT-----TTTT
*  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
                                           +1

      +1
met leu leu gln ala phe leu phe leu leu
TTTTTCTCTCAGAGAAACAGCAAAACAAAAGCTTTTCTTTCACTAACGTATATG ATG CTT TTG CAA GCT TTC CTT TTC CTT TTG
*****  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
CTTTTTCTCAGTGAAACAGCAAAACAAAAGCTTTTCTTTCACTAACGTATATG ATG CTT TTG CAA GCT TTC ATT TTC CTT TTG
*  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
                                           +1

      +1
met leu leu gln ala phe asn phe leu leu

      100
ala gly phe ala ala lys ile ser ala ser met thr asn glu thr ser asp arg pro leu val his phe thr pro
GCT GGT TTT GCA GCC AAA ATA TCT GCA TCA ATG ACA AAC GAA ACT AGC GAT AGA CCT TTG GTC CAC TTC ACA CCC
*  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
GCT GGC TTT GCA GCT AAG ATA TCA GCA TTA ATG ACA AAC GAA ACT AGC GAT AGA CCT TTG GTC CAC TTC ACC CCC
ala gly phe ala ala lys ile ser ala Leu met thr asn glu thr ser asp arg pro leu val his phe thr pro
*  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
                                           100

      100
asn lys gly trp met asn asp pro asn gly leu trp tyr asp glu lys asp ala lys trp his leu tyr phe gln
AAC AAG GGC TGG ATG AAT GAC CCA AAT GGC TTG TGG TAC GAT GAA AAA GAT GCC AAA TGG CAT CTG TAC TTT CAA
*  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
AAC AAG GGT TGG ATG AAT GAT CCA AAT GGA TTG TGG TAC GAT GCA AAA GAA GGT AAA TGG CAT CTC TAC TTT CAG
asn lys gly trp met asn asp pro asn gly leu trp tyr asp ala lys glu gly lys trp his leu tyr phe gln
*  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *

      200
tyr asn pro asn asp thr val trp gly thr pro leu phe trp SUC2
TAC AAC CCA AAT GAC ACC GTA TGG GGT ACG CCA TTG TTT TGG *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
*  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
TAC AAT CCA AAT GAC ACT GTT TGG GGT TTG CCA CTC TTC TGG *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
tyr asn pro asn asp thr val trp gly Leu pro leu phe trp SUC7
*  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
                                           200

```

secreted invertase (8) is conserved at SUC7, as is the surrounding region. However, position -40 is 10 nucleotides closer to the TATA box in SUC7 than SUC2, and therefore either the 5' endpoint of the mRNA or the spacing between the TATA box and the 5' end is not conserved.

Farther upstream, the region between -259 and -399 is less well conserved; the overall homology among the matched residues is only about 50%, and there are not even any strikingly impressive small patches of homology. As mentioned above, this region was previously shown to contain no essential sequences (9). Yet farther upstream lies a second region of significant sequence conservation between SUC2 and SUC7 which corresponds to the upstream regulatory region.

Conservation of upstream regulatory region

Previous studies of the SUC2 gene identified a region between -418 and -650, relative to the translational start site, required for derepression of secreted invertase; the most crucial sequences appeared to lie downstream of -496 (9). This upstream region was shown to confer regulated expression to a heterologous gene not normally regulated by glucose repression (a LEU2-lacZ fusion) and to mediate the regulatory effects of trans-acting genes (10). The upstream region is conserved in the SUC7 5' noncoding sequence. The homology between the SUC2 and SUC7 sequences is over 70% between positions -400 and -508, and in part of this region, between positions -400 and -451, the homology is about 80%. Between -509 and -624 the overall homology drops to 48%, although this region includes a perfectly homologous 10-nucleotide sequence around position -600. No

Figure 2. Comparison of the nucleotide sequences of the 5' ends of the SUC2 and SUC7 genes. The SUC2 sequence, shown on the upper lines, is taken from refs. 9 and 11, except for the sequence between -784 and -637. The sequences of the mRNA-identical strands are presented. Asterisks between nucleotides in the two sequences indicate identical nucleotides. The open reading frame encoded by the 1.9-kb SUC2 mRNA and the corresponding open reading frame of SUC7 are translated into amino acids, and differences between the two polypeptides are boxed. The SUC2 TATA box and its SUC7 counterpart are overlined. The position of the 5' end of the 1.9-kb SUC2 mRNA (8) is marked with an asterisk above the sequence. The site for cleavage of the signal peptide (8, 28) is indicated by a vertical arrow.

```

                                1600
val asn met thr thr gly val asp asn leu phe tyr ile asp lys phe gln val arg glu val lys ter
GTG AAC ATG ACC ACT GGT GTC GAT AAT TTG TTC TAC ATT GAC AAG TTC CAA GTA AGG GAA GTA AAA TAGAGGTTATAA
*** **
GTG AAC ATG ACC ACT GGT GTC GAT AAT TTG TTC TAC ATT GAC AAG TTC CAA GTA AGG GAA GTA AAA TAGAGGTCATAA
val asn met thr thr gly val asp asn leu phe tyr ile asp lys phe gln val arg glu val lys ter
                                1600

                                1700
AACTTATTGCTCTTTTATTTTTCAAAAGCCATTCTAAAGGGCTTTAGC--AACGAGTGACGAATGTAACCTTTATGATTTCAAAGAATAACCTCC
*****
AACTTATTGCTCTTTTATTTTTCAAAAGCCATTCTAAAGGGCTTCGTAATAACGAGTGACGAATGTAACCTTTATGATTTCAAAGAATAACCTCC
                                1700

                                1800
AAACCATGAAAAATGT-----ATTTTATTTTATTTTCTCCGACCCAGTTACCTGGAATTGTTCTTTATGTAAGTATATATAAGTATAATTCTCTTA
*****
AAACCATGAAAAATTTATTTTATTTTATTTTCTCCGACCCAGTTACCTGGAATTGTTCTTTATGTAAGTATATATAAGTATAAGTATGTTT
                                1800

                                1900
AAAATTTTACTACTTTGCAATAGACATCATTTTTCACGTAATAAACCCACAATCGTAATGTAGTTGCCTTACACTACTAGGATGGACCTTTTTCGCTT
* * * * *
AGACTTGAAAAATGTGTCATTTTCTTATTTTGATATTATGAGTTGCTCATTCTCTGCTGTGAGACTTTTGTCTCTTCAAAGGAACGAACAAGTGA
                                1900

                                2000
TATCTGTTTGTACTGACACAATGAAACCGGGTAAAGTATTAGTTATGTGAAAAATTTAAAGCATTAAAGTAGAAGTATACCATATTGTAAAAA
* * * * *
AAAAAATGTACGGTAGGTAGACATAGCCCATCATATAACAAGACTGCATACTAAAAACATATTATTAGGCCATGCTATTGAGAGAAGCTAAGGAT
                                2000

AGCGTTGCTTCTACGTAAAGTGTTCTCAAAAAGAAGTAGTGAGGGAAATGGATACCAAGCTATCTGTGAACAGGAGCTAAAA SUC2
* * * * *
ACATGTAAGTACGCTAGCTCTGCATTTTAAATTATTTTCGGAGGTTATAAGAAAGACAGACAGTGCGAAGAGACCGCAATTTT SUC7

```

Figure 3. Comparison of the nucleotide sequences of the 3' ends of SUC2 and SUC7. The SUC2 sequence, shown on the upper lines, is taken from ref. 11. The sequences of the mRNA-identical strands are presented. Asterisks are placed between identical nucleotides in the two sequences. The open reading frames are translated into amino acids. The two termination codons immediately following each coding sequence are marked. Possible recognition sequences for transcriptional termination or poly(A) addition (see text) are overlined.

significant homology was detected upstream of -625. The position of the conserved upstream regulatory region relative to the TATA box is approximately the same at both SUC loci; the 3' end of the region is 35 nucleotides closer to the TATA box at SUC7.

3' noncoding region

At the 3' end of the genes, the homology is nearly perfect (98%) from the termination codon at +1597 to position +1791 (Figure 3). The homology then ends abruptly, and the sequences from +1792 through +2083 show no significant homology. The end of the homology coincides approximately with the 3' end of the SUC2 mRNA, which maps near +1817 (7, 11). Sequences that have been

implicated as signals for transcription termination and polyadenylation in yeast are included within the conserved region. The sequence TATGT, found at position +1771, is often located 10 to 40 bp before the poly(A) site (20). the sequence TATATAAG at position +1780 resembles the consensus sequence TAAATAAA/G, which is often found 25 to 40 nucleotides before the poly(A) addition site in yeast genes (21). In addition, the sequence TTTTAT is repeated near position +1730 in both genes; this sequence resembles the octamer TTTTATA that has been implicated in transcriptional termination in yeast (22).

Functional comparison of SUC2 and SUC7

Because of the conservation of the coding and regulatory regions between SUC2 and SUC7, it seemed likely that the two genes would exhibit a high degree of functional similarity. To examine the expression of the SUC7 gene, we assayed the production of the two forms of invertase under conditions of glucose repression and derepression. First, an extract prepared from cells carrying SUC7, and no other active SUC⁺ gene, was subjected to electrophoresis on a nondenaturing polyacrylamide gel, which was then stained in situ for invertase activity (23). Two forms of invertase were detected that migrated close to the positions of the enzymes encoded by SUC2, although the nonglycosylated SUC7 invertase showed slightly lower mobility than the nonglycosylated SUC2 species (data not shown). The regulation of SUC7 expression appeared identical to that of SUC2: the nonglycosylated invertase was produced constitutively, and synthesis of the glycosylated enzyme was regulated by glucose repression. The glycosylated invertase is secreted into the periplasmic space, as is the product of the SUC2 gene, because invertase activity can be detected by assaying whole cells (Table 1). These quantitative assays also confirmed that the secreted invertase is regulated by glucose repression. Moreover, previous genetic analysis has indicated that SUC2 and SUC7 are regulated by at least some of the same trans-acting elements; both genes require the SNF1 gene for derepression (17).

To compare quantitatively the levels of expression from SUC2 and SUC7, we constructed congenic strains, so that gene expression would not be differentially affected by genetic

Table 1. Expression of secreted invertase activity from SUC7

<u>Strain</u>	Relevant <u>Genotype</u>	<u>Invertase Activity</u> ^a	
		<u>Repressed</u>	<u>Derepressed</u>
DBY615	<u>SUC7</u> ⁺	<1	45
MCY130 ^b	<u>SUC7</u> ⁺	<1	27
MCY257 ^b	<u>SUC7</u> ⁺	<1	13
MCY1207 ^b	<u>SUC7</u> ⁺	<1	34
MCY501 ^c	<u>SUC2</u> ⁺	<1	200

^aUnits are micromoles of glucose released per minute per 100 mg (dry weight) of cells

^bCongenic to S288C

^cIsogenic to S288C

background, and assayed secreted invertase activity. The SUC7 gene produced about ten fold less invertase activity than did SUC2 in backgrounds congenic to S288C, the strain from which our SUC2 gene is derived (Table 1). Interestingly, the SUC7 gene showed the highest expression in DBY615, which is the strain in which SUC7 was originally identified.

To determine whether the difference in the level of invertase expression from the SUC2 and SUC7 genes resulted from a difference in the expression of mRNAs, we examined the mRNAs in glucose-repressed and derepressed SUC2 and SUC7 strains by gel-transfer hybridization (24). A SUC7 strain (MCY1207) carrying a deletion of the SUC2 gene was used because the naturally occurring suc2^o allele is a mutant gene that encodes 1.8- and 1.9-kb mRNAs (7). Poly(A)-containing RNAs were separated by gel electrophoresis, transferred to nitrocellulose (25) and hybridized to a probe homologous to the coding region of SUC genes. Figure 4 shows that the expected 1.8-kb SUC2 mRNA was detected in poly(A)-containing RNA from glucose-repressed SUC2 cells and that the 1.9-kb mRNA was detected in derepressed cells. A glucose-repressible 1.9-kb mRNA was detected in SUC7 cells, but the derepressed level was lower than that of the 1.9-kb SUC2 mRNA. No 1.8-kb SUC7 mRNA was detected in repressed cells. To confirm that comparable amounts of RNA were loaded in all lanes,

Figure 4. Gel-transfer hybridization analysis of *SUC2* and *SUC7* mRNAs. Poly(A)-containing mRNAs were prepared from glucose repressed (R) and derepressed (D) cultures of MCY1207 (*SUC7 suc2-Δ9*) and MCY501 (*SUC2*). Cultures were grown as described in Methods except that derepression was for 1 h. RNAs were isolated by a modification (G. Stettler, pers. comm.) of the method of Tuite et al. (29). Yeast cells were vortexed with glass beads in 0.1 M Tris-HCl (pH 7.5), 0.1 M LiCl, 1 mM EDTA, 0.5% SDS, 1% diethylpyrocarbonate. The lysate was treated with 400 μg/ml proteinase K for 30 min at 37° and extracted with phenol:chloroform:isoamylalcohol. Nucleic acids were recovered by precipitation with ethanol, and poly(A)-containing RNAs were selected by chromatography on oligo(dT) - cellulose (30). RNAs were separated by electrophoresis on a 1.5% agarose gel containing formaldehyde (12) and transferred to nitrocellulose (25). ³²P-labelled probes were prepared (31) from plasmids pRB117 (7), which carries the 5' portion of the *SUC2* coding region, and YEp351 (gift of J. Hill and T.J. Koerner), which carries the *LEU2* gene. RNAs homologous to the probes were detected by hybridization (3) and autoradiography.

the filter was hybridized with a probe specific to the *LEU2* gene, which is not glucose-repressible (26); the amount of the 1.3-kb *LEU2* transcript (27) detected in each RNA sample was nearly identical. The same relative amounts of the *SUC* mRNAs were detected in three experiments. These results indicate that at least part of the difference in the level of secreted invertase expression from *SUC2* and *SUC7* can be attributed to a difference in the level of stable 1.9-kb mRNA. Our failure to detect a 1.8-kb *SUC7* mRNA in glucose-repressed cells is consistent with the observation that the activity of the nonglycosylated invertase detected by the gel assay was considerably lower in

repressed SUC7 (MCY1207) cultures than in repressed SUC2 cultures (data not shown).

DISCUSSION

We have examined the relationship between SUC2 and one of the telomeric SUC genes of Saccharomyces. Previous studies had shown that the dispersed SUC gene family includes the SUC2 gene, which is the gene most commonly found in Saccharomyces strains, and five SUC genes that are located near telomeres and share extensive homologous flanking sequences (SUC1, SUC3, SUC4, SUC5 and SUC7). The evidence suggested that the dispersal of the telomeric SUC genes to different chromosomes occurred by rearrangements of chromosome termini, probably as a result of homologous recombination between repeated sequences common to many telomeres (3). The evolutionary relationship of SUC2 and the telomeric genes, however, remained obscure. In this paper we have compared the nucleotide sequences of SUC2 and a representative of the telomeric group, SUC7. The homology between the two includes the coding region, 5' noncoding sequences extending as far as 624 nucleotides upstream, and 3' noncoding sequences extending approximately to the end of the mRNA. No significant homology was detected upstream from -625. Thus, it appears that a duplication of the sequence between roughly -625 and +1800 occurred in the yeast genome.

The mechanism by which this duplication was effected remains unclear. There is no evidence either suggesting or excluding the involvement of reverse transcription in this process; the SUC genes do not have introns and there is no residual poly(dA:dT) tract at the 3' end of either gene. Nor is there evidence suggesting a transposon-mediated event; the SUC2 gene is flanked by unique sequences for many kilobases on either side (3). In the absence of any mechanistic evidence, it is possible that either SUC2 or a telomeric SUC gene could have been the progenitor of the family. However, the fact that most, if not all, strains carry SUC gene sequences at the SUC2 locus (either SUC2⁺ or a defective allele), while many strains do not carry SUC genes at any of the telomeric loci (1), suggests that the family evolved from the SUC2 gene.

The level of homology between SUC2 and SUC7 varies in different regions of the duplicated sequence. A high degree of conservation was found in the amino-terminal and carboxyl-terminal coding regions, including the signal sequence. Similar conservation was found in the 3' noncoding region, and putative signals for transcriptional termination and polyadenylation were conserved. Interestingly, the homology in the 3' noncoding sequence extends to the approximate end of the mRNA and then terminates abruptly. This abrupt endpoint may reflect the molecular events leading to gene duplication or may be a consequence of selection for conservation only of functional regions of sequence.

The 5' noncoding region is highly conserved between position -140, near the TATA box, and the translational start site; this region presumably contains promoter elements and recognition signals for translational initiation. The sequence surrounding the nucleotide that corresponds to the 5' end of the 1.9-kb SUC2 mRNA is perfectly conserved, although the spacing between this sequence and the TATA box is not. The TATA box itself is not perfectly conserved.

The 5' noncoding region also includes a second region of strong homology located upstream from nucleotide -400. The SUC2 upstream regulatory region, which has been shown to be essential for expression of secreted invertase (9), has a highly conserved counterpart at the SUC7 locus. With the exception of two short gaps in the homology, the sequence between -400 and -508 is more than 70% homologous. It is noteworthy that although the homology between SUC2 and SUC7 is only 48% in the region between -624 and -509, a 10-nucleotide sequence at -600 is perfectly conserved. Further studies have suggested a functional role for a similar sequence as an enhancer element (L.S. and M.C., in preparation).

An unusual feature of this gene duplication is that two highly conserved regions 5' to the coding sequence are separated by a region of significantly lower homology. The homology between SUC2 and SUC7 is considerably less impressive in the region extending from the TATA box to the upstream regulatory region; this less conserved region has been shown by deletion analysis to be dispensable for expression (9). These findings suggest that

the sequence divergence observed within this region resulted from the absence of selective pressure for functionality.

Studies comparing the expression of the SUC2 and SUC7 genes revealed a quantitative difference: the SUC7 gene produces roughly ten fold lower secreted invertase activity than does SUC2 in strains of the same genetic background. This lower level of activity results, at least in part, from a lower level of stable 1.9-kb mRNA in derepressed cells. We suggest that the SUC7 gene may simply be transcribed less efficiently than SUC2 due to differences in the upstream regulatory region and/or the proximal 5' noncoding sequence. Another possibility is that chromosomal position affects expression. Studies to identify sequence elements that are important for regulation and for expression are in progress.

ACKNOWLEDGMENTS

We thank Steve Goff and Janie Weiss for help with the computer analysis, Ron Taussig for expert technical assistance and Lenore Neigeborn for performing invertase assays.

This work was supported by Public Health Service grant GM32065 from the National Institutes of Health. L.S. was a predoctoral trainee supported by grant GM07088 from the National Institutes of Health.

*To whom correspondence should be addressed

REFERENCES

1. Carlson, M. and Botstein, D. (1983) *Mol. Cell. Biol.* 3, 351-359.
2. Carlson, M., Osmond, B.C. and Botstein, D. (1980) *Cold Spring Harbor Symp. Quant. Biol.* 45, 799-803.
3. Carlson, M., Celenza, J.L. and Eng, F.J., *Mol. Cell. Biol.*, in press.
4. Chan, C.S.M. and Tye, B.-K. (1983) *Cell* 33, 563-573.
5. Mortimer, R.K. and Schild, D. (1982) in *The Molecular Biology of the Yeast Saccharomyces. Metabolism and Gene Expression*, Strathern, J.N., Jones, E.W. and Broach, J.R. Eds., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
6. Carlson, M., Osmond, B.C. and Botstein, D. (1981) *Genetics* 98, 41-54.
7. Carlson, M. and Botstein, D. (1982) *Cell* 28 145-154.
8. Carlson, M., Taussig, R., Kustu, S. and Botstein, D. (1983) *Mol. Cell. Biol.* 3, 439-447.

-
9. Sarokin, L. and Carlson, M. (1984) *Mol. Cell. Biol.* 4, 2750-2757.
 10. Sarokin, L. and Carlson, M. (1985) *Mol. Cell. Biol.*, in press.
 11. Taussig, R. and Carlson, M. (1983) *Nucleic Acids Res.* 11, 1943-1954.
 12. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) *Molecular Cloning, A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
 13. Maxam, A.M. and Gilbert, W. (1980) *Methods Enzymol.* 65, 499-560.
 14. Dayhoff, M.O., Barker, W.C. and Hunt, L.T. (1983) *Methods Enzymol.* 91, 524-545.
 15. Celenza, J.L. and Carlson, M. (1984) *Mol. Cell. Biol.* 4, 49-53.
 16. Goldstein, A. and Lampen, J.O. (1975) *Methods Enzymol.* 42C, 504-511.
 17. Carlson, M., Osmond, B.C. and Botstein, D. (1981) *Genetics* 98, 25-40.
 18. Hubbard, S.C. and Ivatt, R.J. (1981) *Ann. Rev. Biochem.* 50, 555-583.
 19. Struck, D.K. and Lennarz, W.J. (1980) in *The Biochemistry of Glycoproteins and Proteoglycans*, Lennarz, W.J. Ed. pp. 35-83, Plenum Press, New York.
 20. Zaret, K.S. and Sherman, F. (1982) *Cell* 28, 563-573.
 21. Bennetzen, J.L. and Hall, B.D. (1982) *J. Biol. Chem.* 257, 3018-3025.
 22. Henikoff, S., Kelly, J.D. and Cohen, E.H. (1983) *Cell* 33, 607-614.
 23. Gabriel, O. and Wang, S.-F. (1969) *Anal. Biochem.* 27, 545-554.
 24. Alwine, J.C., Kemp, D.J. and Stark, G.R. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5350-5354.
 25. Thomas, P.S. (1980) *Proc. Natl. Acad. Sci. USA* 77, 5201-5205.
 26. Martinez-Arias, A., Yost, H.J. and Casadaban, M.J. (1984) *Nature* 307, 740-742.
 27. Andreadis, A., Hsu, Y.-P., Hermodson, M., Kohlhaw, G. and Schimmel, P. (1984) *J. Biol. Chem.* 259, 8059-8062.
 28. Perlman, D., Halvorson, H.O. and Cannon, L.E. (1982) *Proc. Natl. Acad. Sci. USA* 79, 781-785.
 29. Tuite, M.F., Plesset, J., Moldave, K. and McLaughlin, C.S. (1980) *J. Biol. Chem.* 255, 8761-8766.
 30. Aviv, H. and Leder, P. (1972) *Proc. Natl. Acad. Sci. USA* 69, 1408-1412.
 31. Rigby, P.W.J., Dieckmann, M., Rhodes, C. and Berg, P. (1977) *J. Mol. Biol.* 113, 237-251.
-