Volume 14 Number 16 1986 Nucleic Acids Research

Nucleotide sequence of the yeast cell division cycle start genes *CDC28, CDC36, CDC37,* and *CDC39,* and a structural analysis of the predicted products

Jill Ferguson+, Jeong-Yau Ho+, Thomas A.Peterson* and Steven I.Reed+

Biochemistry and Molecular Biology Section, Department of Biological Sciences, University of California, Santa Barbara, CA 93106, USA

Received 14 April 1986; Revised 30 June 1986; Accepted 7 July 1986

ABSTRACT

The nucleotide sequences of the yeast cell division cycle start genes CDC36, CDC37, and CDC39 are presented. An open reading frame corresponding in size and mapped position to the mRNA for each gene was revealed. These sequences, as well as that of the CDC28 gene, were analyzed for the presence of consensus sequences postulated to be transcriptional or translational signals, or to be involved in mRNA processing. In addition, the predicted protein products of the four genes were subjected to a number of structural and statistical analyses including codon usage bias analysis, secondary structure analysis and hydropathicity analysis.

INTRODUCTION

In the budding yeast, Saccharomyces cerevisiae, it has been possible to isolate cell division cycle (cdc) mutations which disrupt the cell division cycle in a stage-specific manner (1-4). Studies of cdc mutants have led to the formulation of a model where the events of cell division are organized into a number of interrelated dependent pathways (2,4-6). Because all pathways diverge from G1, cell division can, in theory, be controlled at a single point within this interval. Yeast cells deprived of an essential nutrient (7,8) or responding to mating pheromone (9) cease from further division cycles and become synchronous at a point in G1 which has been designated "start" (2). "Start" is thought to be the stage, therefore, where cell division is controlled. This report is part of an ongoing genetic and molecular analysis of "start" with the aim of elucidating division control at the molecular level.

cdc mutations in four unlinked complementation groups, CDC28, CDC36, CDC37, and CDC39, arrest cells at start (2,3). It is therefore likely that the products of the genes defined by these complementation groups play a role in the division control process. For this reason, the genes were cloned (10,11) and then subjected to DNA sequence analysis. The DNA sequence of one of these start genes, CDC28, has been reported (12). The inferred amino acid

sequence of the CDC28 gene product is homologous to several vertebrate protein kinases, including those encoded by members of the src oncogene family (12). Furthermore the CDC28 gene product has been shown to have a protein kinase activity in an immune complex assay (13). These results suggest that protein phosphorylation may play an important role in control of cell division in yeast. The isolation and transcriptional characterization of three other start genes CDC36, CDC37, and CDC39 have previously been reported (11). The inferred amino acid sequence of the CDC36 gene has been reported to have significant homology to the CDC4 product, also required early in the yeast cell division cycle, and to the product of the transformation-specific sequence of avian erythroblastosis virus E26, ets (14). In virus E26, the ets sequence is linked in frame with Δgag and myb^E in the tripartite structure 5'-Δgag-myb^E-ets-3', comprising the E26 transforming oncogene (14). The CDC36 gene product has homology to the 149 residue C-terminal portion of the CDC4 gene product, and to a 206 residue region from the middle of the ets product (14). These data suggest a possible relationship between genes involved in yeast division control and vertebrate proto-oncogenes.

We report here the nucleotide sequence of the CDC36, CDC37, and CDC39 genes, and an analysis of the 5' and 3' untranslated regions of these, as well as of the CDC28 gene, for consensus sequences postulated to be transcriptional or translational signals, or to be involved in mRNA processing. Secondary structures for the inferred amino acid sequences of the four genes under study have been predicted according to the method of Chou and Fasman (15,16), and protein hydropathicities (an estimate of hydrophobicity) determined by the method of Kyte and Doolittle (17). The codon usage biases (18) for these genes have been examined in relation to the abundancies of the corresponding mRNAs in the cell.

MATERIALS AND METHODS

DNA Sequence Analysis

The CDC36, CDC37, and CDC39 genes were sequenced by the M13-dideoxy method of Sanger et al. (19) used in conjunction with M13 phages (20). DNA deletions were obtained either by progressive digestion with Ba131 nuclease (12), or the sequential action of Exonuclease III from E. coli and S1 nuclease from Aspergillus (21), and sequenced directly. When Ba131 was used, deletions were initiated at a centrally located unique restriction site and deletion end points were juxtaposed to the sequencing primer by digestion

with a restriction enzyme which cleaves at a proximal site in the cloning polylinker segment followed by religation (12). Ba131 digestions were done at 30° using 1 unit Ba131 (BRL) per µg of DNA at a concentration of .02 µg DNA/µl of reaction mix, in 600 mM NaCl, 20 mM Tris-HCl (pH 8), 12 mM $CaCl_2$, 12 mM $MgCl_2$, and 1 mM EDTA. These conditions resulted in a degradation rate of approximately 100 bp per end per minute. Depending on the length of the DNA sequence of interest and the extent of digestion desired, in some cases several time points of Ba131 digestion were combined into a single pool. Each pool of DNA was extracted with phenol, precipitated by addition of ethanol, and digested with an appropriate restriction enzyme to cleave in the cloning polylinker segment, in order to remove the residual DNA fragment between the sequencing primer and the site of initiation of Ba131 digestion. This sequencing method requires having a "buffer" region of dispensable DNA in this region. A 5 minute incubation at 65° followed. Each pool of DNA was then treated for 10 minutes at room temperature with 4 units of E. coli DNA Polymerase I (large fragment; BRL) per µg DNA at a concentration of .05 µg DNA/µl reaction mix in 13 mM each of Tris-HCl (pH 7.5), NaCl, $MgCl_2$, and DTT, and 0.1 mM each of deoxynucleotide triphosphate, followed by extraction with phenol and precipitation by addition of ethanol. Ligations were performed at room temperature for 6-18 hours with 2 units T4 DNA Ligase (BRL) per µg DNA at a concentration of .04 µg DNA/µl reaction mix in 1 mM ATP, 66 mM Tris-HCl (pH 7.5), 6.6 mM $MgCl_2$, 10 mM DTT, and 100 µg/ml bovine serum albumin. Each DNA pool was then transfected into competent JM103 or JM101 cells (20). Deletion end points were aligned by means of restriction mapping so that appropriate phages could be chosen for sequencing. The procedure employed to produce random deletions for sequencing using the sequential action of Exonuclease III and S1 nuclease was that of Henikoff (21). We found that for this procedure it is mandatory to use M13 replicative form DNA prepared by a method which incorporates an alkaline lysis step, such as that of Ish-Horowicz and Burke (22). DNA prepared otherwise is labile to Exonuclease III digestion throughout the molecule and cannot be kept intact after Exonuclease III treatment. All DNA sequences were determined more than once, either on separate strands or from different endpoints on the same strand. The authors accept responsibility for the accuracy of the DNA sequences reported.

Computing

The computer analyses were performed using the PEP computer program available through Bionet. The PEP secondary structure analyses are based on the algorithm of Chou and Fasman (15,16). The ambiguities in the computer-

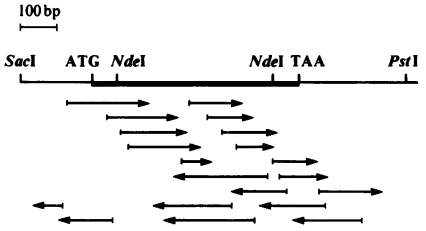


Fig. 1 Sequencing strategy for the CDC36 gene. The directions and limits of the sequencing runs are shown below a partial restriction map of the region. The open reading frame found by sequence analysis is indicated by a thick line, with the initiating ATG and terminating TAA codons labelled.

predicted structures were resolved using the rules given in Chou and Fasman (16). The PEP hydropathicity analyses are based on the algorithm of Kyte and Doolittle (17), which uses a moving-segment approach that continuously determines the average hydropathy within a segment of predetermined length as it advances through the sequence. For globular proteins there is a correspondence between the interior and exterior portions, and the predicted hydrophobic and hydrophilic portions of the protein (17).

RESULTS AND DISCUSSION

We report here the nucleotide sequences of three yeast cell division cycle start genes, CDC36, CDC37, and CDC39. The genes were isolated on recombinant plasmids, and their identities confirmed by their ability to recombine at their respective genomic loci (11). The genes were further localized on the plasmids by subcloning, RNA blot analysis, and R-loop

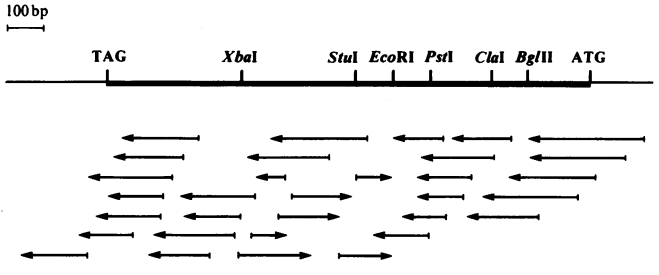


Fig. 2 Sequencing strategy for the CDC37 gene. The directions and limits of the sequencing runs are shown below a partial restriction map of the region. The open reading frame found by sequence analysis is indicated by a thick line, with the initiating ATG and terminating TAG codons labelled.

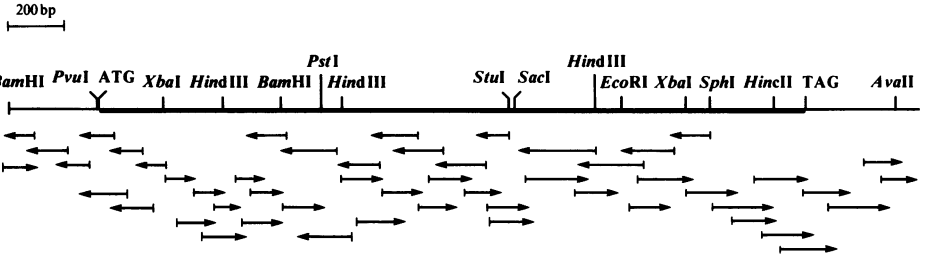


Fig. 3 Sequencing strategy for the CDC39 gene. The directions and limits of the sequencing runs are shown below a partial restriction map of the region. The open reading frame found by sequence analysis is indicated by a thick line, with the initiating ATG and terminating TAG codons labelled.

```
                                                -200                                                -150
                                                CCTCTATGAGTCTGTGTATGATTTATAAAGAGTGAGCTCTTTTGTTATGAAGTATATTATTATGTGAGTATTTTTTTATAACTGCG

                                    -100                                                -50                                                -1
AAAAAGCGAGGCGAACATTAATAAGATAATAGCAAAACAGGCATTTTTTAAACTAGATATGCAAAAACTACAAGGAAAAACAGGAAAAGAAGAGAATCAAAATATACTATTGCCTACATA

1                                                   50                                                  100
ATGGAAAAATTTGGTTTAAAAGCGCTAGTACCGCTCCTTAAGCTAGAGGACAAAGAACTTTCAAGTACATATGACCATTCCATGACCTTAGGAGCGGACTTATCGTCGATGCTATATTCG
 M  E  K  F  G  L  K  A  L  V  P  L  L  K  L  E  D  K  E  L  S  S  T  Y  D  H  S  M  T  L  G  A  D  L  S  S  M  L  Y  S

                                    150                                                 200
TTGGGTATTCCAAGAGATTCACAAGATCATAGAGTCCTCGATACTTTTCAATCACCGTGGGCAGAAACATCCAGAAGCGAAGTGGAGCCCCGATTTTTCACACCAGAATCATTCACCAAT
 L  G  I  P  R  D  S  Q  D  H  R  V  L  D  T  F  Q  S  P  W  A  E  T  S  R  S  E  V  E  P  R  F  F  T  P  E  S  F  T  N

      250                                                 300                                                 350
ATTCCGGGCGTATTACAATCCACTGTAACTCCACCATGCTTCAATTCCATTCAAAATGACCAACAGCGTGTTGCCCTTTTTCAAGATGAAACTCTATTTTTCCTCTTCTATAAACACCCT
 I  P  G  V  L  Q  S  T  V  T  P  P  C  F  N  S  I  Q  N  D  Q  Q  R  V  A  L  F  Q  D  E  T  L  F  F  L  F  Y  K  H  P

                                    400                                                 450
GGTACAGTCATCCAGGAACTAACGTATTTGGAACTCAGGAAAAGGAACTGGAGATACCACAAGACGTTGAAGGCCTGGCTCACCAAAGATCCTATGATGGAACCTATTGTATCCGCTGAT
 G  T  V  I  Q  E  L  T  Y  L  E  L  R  K  R  N  W  R  Y  H  K  T  L  K  A  W  L  T  K  D  P  M  M  E  P  I  V  S  A  D

                                    500                                                 550                                                +1
GGTTTAAGTGAAAGGGGATCATATGTGTTTTTTGACCCACAAAGGTGGGAAAAGTGCCAAAGAGATTTCTTATTGTTTTATAATGCTATTATGTAAGGTAGACAGAGAAACGAAAGAAGT
 G  L  S  E  R  G  S  Y  V  F  F  D  P  Q  R  W  E  K  C  Q  R  D  F  L  L  F  Y  N  A  I  M

                                +50                                                 +100
AAATAAATGTCATTATATGTGATGCGGCGTTTTGGAGCTGAACAGAACGCCCTCTTTTTTTTTTTTTAATGATTAATTGAGAAAAATAGTGTGCACATTTAATTTTGCATGTGGGATTGG

  +150                                              +200                                                +250
TTTTCATACAATTCGACGATCTGGTCAGAAGAGGTTGCATCTTCTACGCCCTTGTCTTCTCTTATACGCAAAAATATTGGAAATCTTAAGGATACGCCTTTATCAAAAGTTGCACTACCT

                                +300                                                +350
GCTTTATAAATAGGCGATAGGGAAAGGTCTGCAGTCAAAACTTCAAACAAAGTCGTAGGTTCGAACCATACATCTGGCTCAGCACTCGAATCGAAAACAAACGTGGCTTTTGGTTCCATC

          +400
AATGATTCGTCGGGGTCAGCCTGTCGTGCAAAGCTGCACATT
```

Fig. 4 The nucleotide sequence of the CDC36 region of the yeast genome and the predicted primary structure of the CDC36 product. The sequence is shown as the rightward 5'->3' strand. The nucleotides of the coding region are numbered 1 through 576. The nucleotides of the 5' flanking region are numbered from -1, starting at the first nucleotide preceding the initiating ATG codon, through -206, proceeding away from the gene. The nucleotides of the 3' flanking region are numbered +1, starting at the first nucleotide after the terminating TAA codon, through +426. The TATA sequences upstream from the coding region are underlined, and the CAAG sequence is in a box. An upstream ATG codon is overlined. The truncation point at the Sau3A restriction site at position 439 is shown by an arrow. Downstream from the coding region the transcription termination sequences TATGT/TAGT...TTT are underlined, and the polyadenylation sequence TAAATAAA is in a box.

mapping (11). All three genes were sequenced by the dideoxy method of Sanger et al. (19) used in conjunction with M13 phages (20).

Figures 1, 2, and 3 show the strategy used to sequence, respectively, a 1.2 Kb segment of DNA containing the CDC36 gene, a 1.7 Kb segment containing the CDC37 gene, and a 3.5 Kb segment containing the CDC39 gene. In each case an open reading frame was revealed, corresponding in size and mapped position to the previously identified mRNA (11). For the CDC36, CDC37, and CDC39 genes the open reading frames are, respectively, 191 codons (Fig. 4), 449 codons (Fig. 5), and 834 codons (Fig. 6) long. All of these genes have been subjected to RNA blotting analysis using an rna2 mutant strain (data not shown) to screen for the presence of introns. The rna2 mutation interferes with mRNA processing and causes accumulation of unspliced RNA species (23). Thus the appearance of higher molecular weight RNA species in an rna2 strain indicates that the corresponding gene contains an intron. Our experiments revealed no introns of any substantial size in any of the genes under study. These observations are consistent with conclusions drawn earlier based on R-loop analyses (11), and an examination of the sequences in the present study for the splice consensus now known for S. cerevisiae (24). The nucleotide sequence of a fourth start gene, CDC28, which has already been reported (12), is shown in Fig. 7 and discussed here as well.

The putative CDC36, CDC37, and CDC39 gene products are, respectively, 22 KD, 51 KD, and 94 KD proteins, of which the first is predicted to be neutral and the latter two are predicted to be acidic. The CDC37 gene product has an excess of 18 acidic over basic residues located in the carboxy-terminal third of the protein, whereas the CDC39 gene product has 11 more acidic than basic residues with two very acidic regions between residues 194-265 and 560-577. There are 20 acidic and 2 basic residues between positions 194 and 265, and 12 acidic and 1 basic residue between positions 560 and 577.

Truncated forms of both the CDC36 and CDC37 genes have been isolated which encode shortened versions of the respective proteins fused to a short adventitious open reading frame in pBR322 DNA. This occurred during subcloning of the two genes using partial digestions of larger plasmids with the restriction enzyme Sau3A and selection by complementation of the respective ts mutations. In the CDC36 gene the truncation occurs at a Sau3A restriction site at position 448 (Fig. 4), resulting in a shortened functional version of the protein containing 79% of the wild type protein sequence spliced to a 17

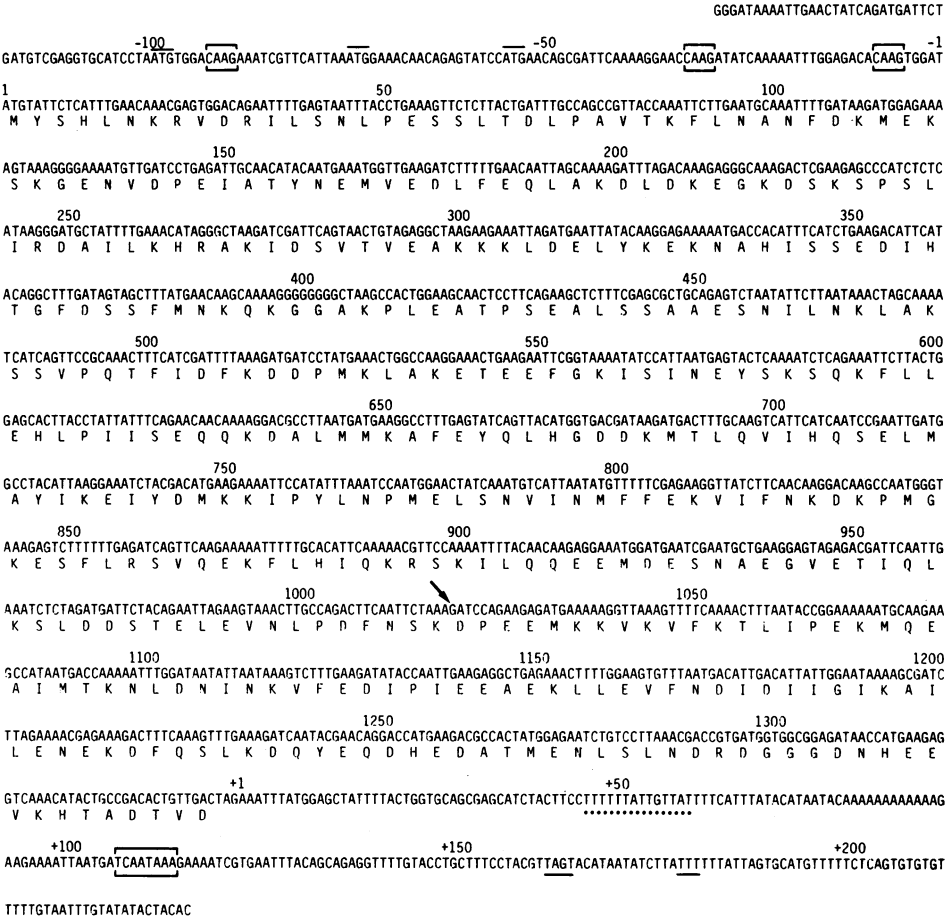
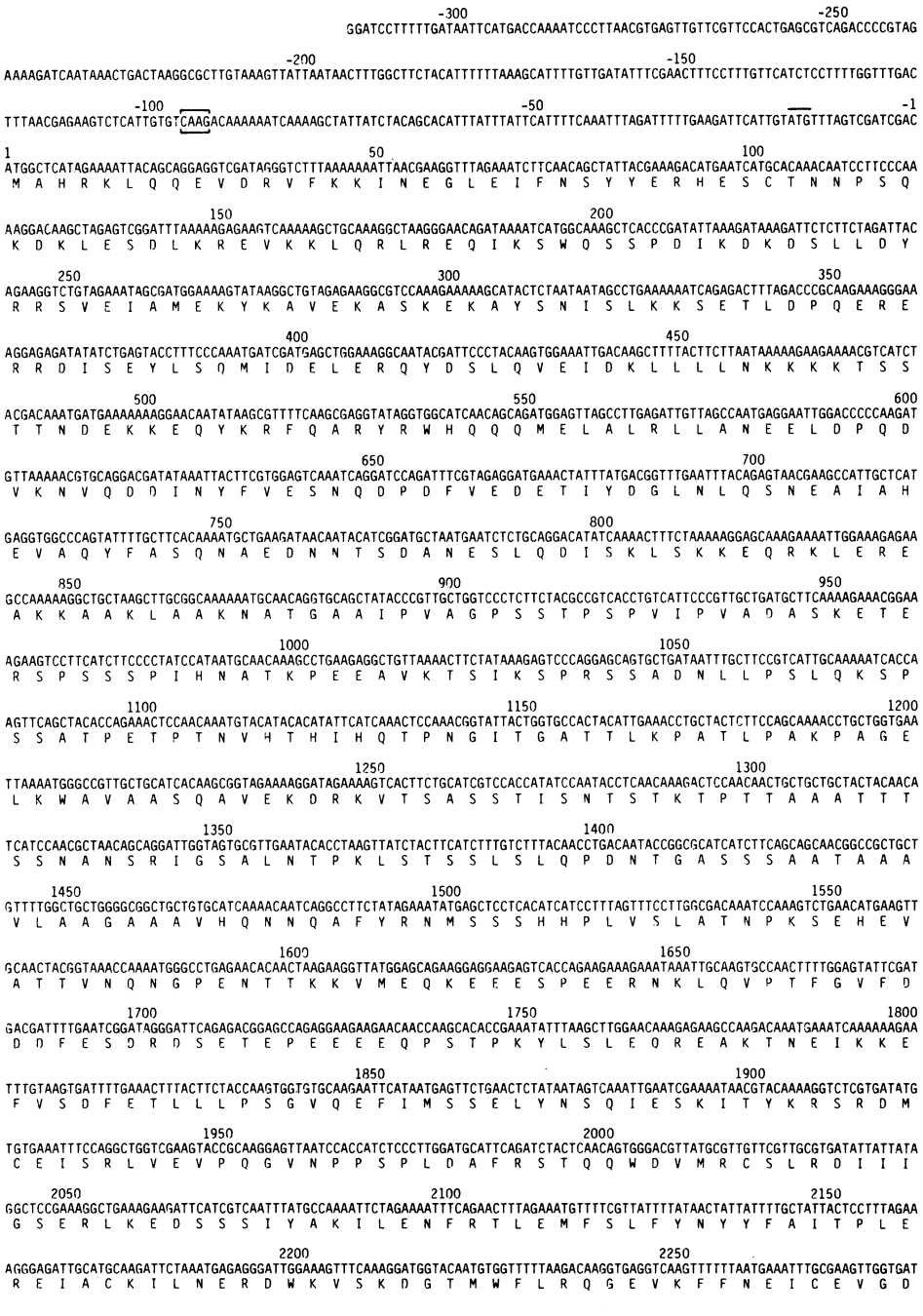


Fig. 5 The nucleotide sequence of the CDC37 region of the yeast genome and the predicted primary structure of the CDC37 product. The sequence is shown as the rightward 5'->3' strand. The nucleotides of the coding region are numbered 1 through 1350. The nucleotides of the 5' flanking region are numbered from -1, starting at the first nucleotide preceding the initiating ATG codon, through -149, proceeding away from the gene. The nucleotides of the 3' flanking region are numbered +1, starting at the first nucleotide after the terminating TAG codon, through +234. The CAAG sequences upstream from the coding region are in boxes. The upstream ATG triplets are overlined. The Sau3A truncation point at position 1018 is shown by an arrow. Downstream from the coding region the transcription termination sequence TAGT...TTT is underlined and the polyadenylation sequence TGAATAAA is in a box. The 14 bp sequence identified by Montgomery et al. (32) is underlined with dots.



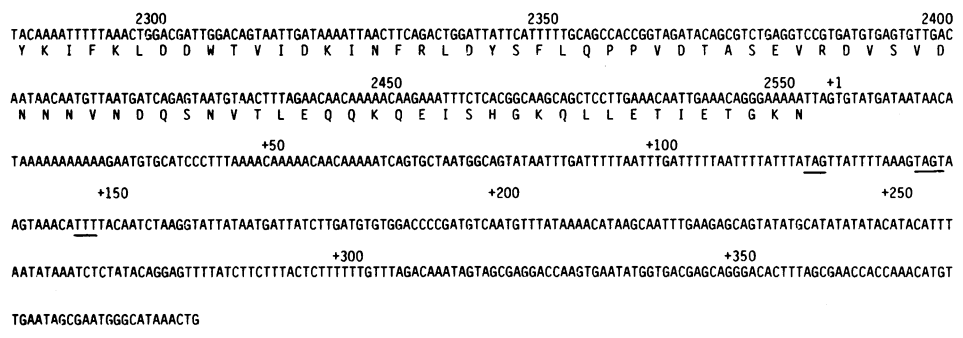


Fig. 6 The nucleotide sequence of the CDC39 region of the yeast genome and the predicted primary structure of the CDC39 product. The sequence is shown as the rightward 5'->3' strand. The nucleotides of the coding region are numbered 1 through 2555. The nucleotides of the 5' flankng region are numbered from -1, starting at the first nucleotide preceding the initiating ATG codon, through -315, proceeding away from the gene. The nucleotides of the 3' flanking region are numbered +1, starting at the first nucleotide after the terminating TAG codon, through +399. The CAAG sequence upstream from the coding region is in a box. The upstream ATG triplet is overlined. Downstream from the coding region the transcription termination sequence TAG...TAGT...TTT is underlined.

codon long open reading frame in pBR322. The truncated form of the CDC37 gene, containing 76% of the wild type sequence, is spliced at the Sau3A site at position 1018 (Fig. 5). The carboxy-terminal segment of the wild type CDC37 gene product that is missing in the truncated version includes most of the acidic portion of the wild type protein, having an excess of 13 acidic over basic residues. In both cases the gene products are functional to the extent that they are able to complement the respective ts allele when on a plasmid containing an ars1 replicator. Such plasmids are normally present at several copies per cell. We do not know for either gene whether the truncated form is capable of complementing the corresponding ts mutation when present at one copy per cell, or whether, at any copy number, a phenotype consistent with the absence of a particular domain of the protein might be observed. For instance, the cdc37-1 ts mutation has a weak kar phenotype in addition to the cell cycle defect (25), and the truncated CDC37 gene might complement one defect but not the other.

We have reported the inferred amino acid sequences of the CDC36, CDC37, and CDC39 genes (Figs. 4, 5, and 6). The CDC28 and CDC36 gene products have previously been reported to be homologous to vertebrate oncogenes (12,14). A similar computer search has to date yielded no proteins homologous to the

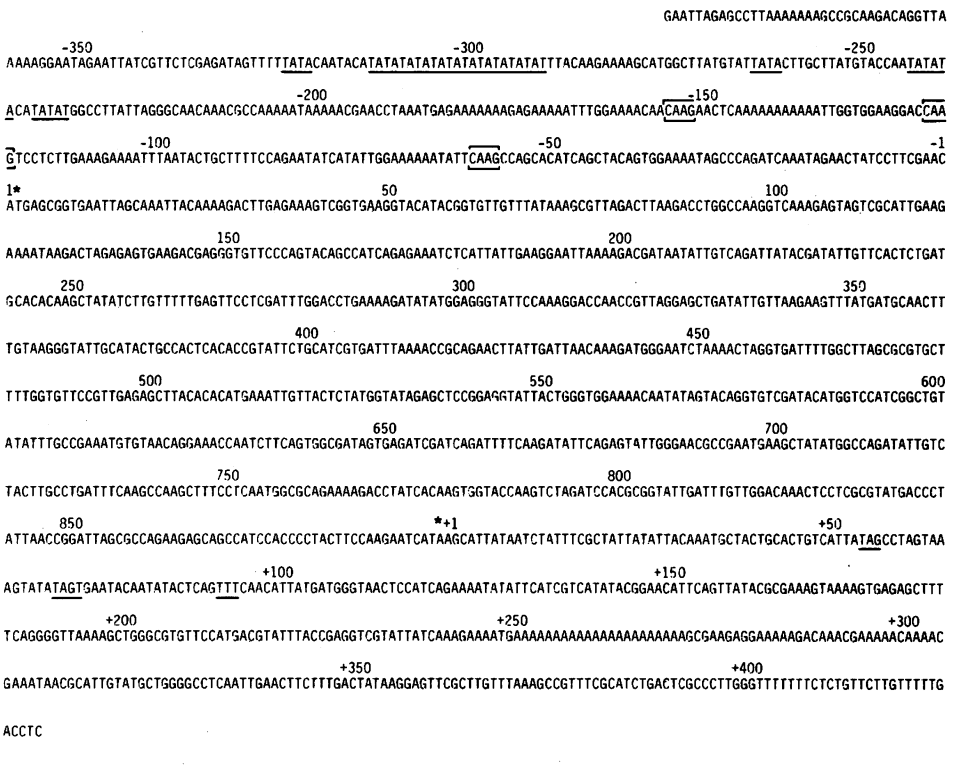


Fig. 7 The nucleotide sequence of the CDC28 region of the yeast genome. The sequence is shown as the rightward 5'->3' strand. The nucleotides of the coding region are numbered 1 through 897. The initiating ATG and terminating TAA codons are marked by asterisks. The nucleotides of the 5' flanking region are numbered from -1, starting at the first nucleotide preceding the initiating ATG codon, through -396, proceeding away from the gene. The nucleotides of the 3' flanking region are numbered +1, starting at the first nucleotide after the terminating TAA codon, through +428. The TATA sequences upstream from the coding region are underlined, and the CAAG sequences are in boxes. Downstream from the coding region the transcription termination sequence TAG...TAGT...TTT is underlined.

CDC37 or CDC39 gene products. Probable secondary structures were predicted for the gene products under study according to the method of Chou and Fasman (15,16). The resulting data (Table 1) indicate that the CDC37 and CDC39 proteins are predicted to have a high percentage of alpha-helical regions when compared with a set of proteins of known structure (15), with the CDC37 protein predicted to have as high a percentage of alpha-helix as any of the set, equivalent to the extremely high alpha helical content of the hemoglobins and myoglobin. Intermediate filament proteins are also thought to

Table 1. Structural Analysis of Predicted Products.

	Percent alpha-helix	Percent beta structure	Hydropathicity
CDC28	42	23	-0.24
CDC36	52	27	-0.40
CDC37	74	12	-0.66
CDC39	63	16	-0.79

have a high alpha-helix content (26). While the implications of the predicted high alpha-helix content, and thus highly organized structure, of the CDC37 and CDC39 proteins must await further study, it is interesting to speculate that they may be structural proteins similar to intermediate filament proteins. This idea is particularly attractive in light of observations that the CDC28 protein appears to be associated with the yeast cytoskeleton (27), and of suggestive genetic evidence that the functions of the CDC37 and CDC28 gene products may be related (27). cdc37/cdc28 double mutants are barely viable at permissive temperature for each of the individual mutations, and may be inviable for some alleles, suggesting that the respective gene products interact.

Hydropathicities have been determined for the CDC28, CDC36, CDC37, and CDC39 gene products according to the method of Kyte and Doolittle (17), which progressively evaluates the hydrophilicity and hydrophobicity of a protein along its amino acid sequence. These data are expressed in Table 1 as the mean hydropathicities averaged over the length of the proteins. The mean hydropathicity of a large set of sequenced soluble proteins has been reported by Kyte and Doolittle (17) to be -0.4. The CDC37 and CDC39 proteins have an average hydropathicity value that is considerably lower than -0.4, indicating that they are more hydrophilic than the average protein. The value for the CDC39 protein is at the lower end of the range reported by Kyte and Doolittle (17) for sequenced soluble proteins, suggesting that it is extremely hydrophilic. On the other hand, the CDC36 protein has an average hydropathicity, and the CDC28 protein is somewhat more hydrophobic than the average of the reported proteins. Membrane-spanning segments of membrane-bound proteins are identified by the Kyte/Doolittle analysis as large uninterrupted hydrophobic

Table 2. Codon Bias Index and Intracellular mRNA Copy Number.

	Codon Bias Index[a]	mRNA copies/haploid cell[b]
CDC28	0.19	7±2
CDC36	0.09	1.5±1
CDC37	0.14	3.1±1.5
CDC39	0.08	4.6±2

[a]Codon Bias Indices were calculated according to Bennetzen and Hall (18).
[b]mRNA copy number values were reported in Breter et al. (11).

segments. It seems clear on the basis of this analysis (data not shown) that none of the four proteins under study is membrane-spanning, as none of them has extended hydrophobic regions typical of such proteins.

Yeast genes show a distinct preference for a subset of the possible coding triplets, the preferred codons tending to be highly homologous to the anticodons of the major yeast tRNA species (18). The degree of bias for the preferred triplets in any gene, called the codon bias index, is a measure of the fraction of codon choices which is biased to 22 preferred triplets. A value of one indicates that for all of the triplets in the mRNA, only codons of the preferred variety are used. A value of zero indicates totally random choice (18). The codon bias index for any gene has been reported to be correlated with its level of expression, with more highly expressed genes being more biased in codon usage than those expressed at lower levels (18). Most explanations of the basis of preferential codon usage involve various aspects of mRNA structure pertaining to translational efficiency, and codon-anticodon binding energies (18). Table 2 shows the codon bias index and intracellular mRNA copy number (11) for the four genes under study. These codon bias indices are among the lowest so far reported. Bennetzen and Hall (18) have reported codon bias indices varying from greater than .90 for the glyceraldehyde-3-phosphate dehydrogenase and alcohol dehydrogenase isozyme 1 genes, both highly expressed, to .15 for the iso-2 cytochrome c gene, which is expressed at a low level. Laughon and Gesteland (28) report a codon bias of less than .10 for the GAL4 gene, which is expressed at about 0.1 mRNA copy per cell (29). All of the values we have reported here are less than .20, with the values for the CDC36 and CDC39 genes being less than .10. These

values correlate well with the low level of expression of these genes, and extend the observed relationship between codon bias index and level of gene expression at the low end of the scale.

The 5' and 3' untranslated regions of the CDC28, CDC36, CDC37, and CDC39 genes have some of the consensus sequences which have been identified in higher eukaryotic genes or in other sequenced yeast genes, and postulated to play a role in transcription, translation, and mRNA polyadenylation. A TATA sequence usually appears about 30 bp upstream from mRNA cap sites in the 5' untranslated region of higher eukaryotic genes (30-32). In the yeast genes characterized so far, TATA sequences are found at variable distances from the transcriptional start sites (33), and some yeast genes appear to have none (34,35). Two of the genes under study, CDC28 and CDC36, have TATA sequences within the sequenced 5' untranslated region. Both of these genes have multiple TATA sequences, and, with the exception of one at position -18 preceding the CDC36 gene which seems too close to the translational start site to be a likely functional TATA box, all are present at a large distance from the translational start site. The five TATA sequences at the 5' end of the CDC28 gene are present at positions -237, -245, -265, -314 and -325, and two distal TATA sequences at the 5' end of the CDC36 gene are present at -130 and -183. The elucidation of the functional role of these distal TATA sequences must await further study. All four start genes are expressed at low levels (11) and it is possible that the TATA sequences preceding the CDC28 and CDC36 genes are not involved in their expression. The CDC37 and CDC39 genes do not have TATA sequences within the 5' untranslated portions so far sequenced, within 149 bp of the CDC37 initiating ATG and 300 bp of the CDC39 initiating ATG.

A hexanucleotide sequence CACACA which has been found close to the initiation codon in several yeast genes (34) is not present in any of the four genes under study. Dobson et al. (34) reported a correlation between highly expressed yeast genes and the presence at the 5' end of the gene of a pyrimidine-rich region followed by the sequence CAAG. It appears that transcription begins at or near CAAG or related sequences in a number of yeast genes, suggesting that this sequence may be involved in transcription initiation or mRNA capping (36). All four of the genes under study have CAAG sequences, of varying numbers and at varying distances from the translational start sites. The CDC36 and CDC39 genes each have a single CAAG sequence, at -50 and -97, respectively, and the CDC28 and CDC37 genes each have three. The CDC28 CAAG sequences are at positions -61, -123, and -156, and the CDC37

CAAG sequences are at positions -9, -33, and -94. The CAAG sequence at position -61 preceding the CDC28 gene is interesting in view of the S1 nuclease mapping of several transcription start sites to this region (Lörincz and Reed, unpublished), and the observation by Burke et al. (36) that CAAG sequences are often present at or near the sites of transcription initiation in yeast genes. In none of the four start genes are any of the CAAG sequences preceded by a pyrimidine-rich region. This may be related to the low level of expression of these genes, however it must be noted in this regard that the yeast GAL4 gene, which is expressed at the extremely low level of 0.1 mRNA molecule per cell (29), does have a canonical CAAG sequence preceded by a CT-rich region at its 5' end (28). Montgomery et al. (35) have noted for several yeast genes a correlation between the presence of a high content of clustered pyrimidine residues preceding the gene, and a high level of its expression. This observation is consistent with the low level of expression of the four start genes and a general lack of pyrimidine-rich tracts in the 5' untranslated regions preceding them.

While eukaryotic ribosomes generally initiate at the first ATG codon encountered from the 5' end of the mRNA molecule, there are exceptions to this rule. Kozak (37) has determined that the nucleotide sequence immediately surrounding the initiating ATG codon does affect the efficacy with which ribosomes initiate at that position. In higher eukaryotes a purine at -3 and a G at +4 appear to constitute a favorable initiator sequence (37). In yeast, an A at -3 is present in virtually all genes so far sequenced (34,36,37), and a pyrimidine, usually a T, at +6 is the most common configuration in efficiently expressed genes (34). The CDC28 and CDC36 genes have the requisite A at the -3 position, while the CDC37 and CDC39 genes have the much less common G residue at that position. These two genes however have the favored T at the +6 position, whereas the CDC28 gene has the alternate pyrimidine C, and the CDC36 gene has a purine at that position. The CDC39 gene is preceded at position -17 by an out-of-phase ATG codon flanked by a pyrimidine at -3 and a T at +6. This ATG has part of the favored environment and almost certainly would act as an initiating ATG for some fraction of the ribosomes scanning a CDC39 mRNA molecule. It is followed by a short open reading frame of 26 codons. The presence of this upstream ATG might be expected to contribute to the low level of expression of the CDC39 gene. Three upstream ATG triplets precede the CDC37 initiating ATG at positions -56, -76, and -101, and one precedes the CDC36 initiating ATG at -62. None of these is preceded at -3 by an A or followed at position +6 by a T, nor is

it known which, if any, of the triplets is present within the transcribed region. None would be expected to act as an efficient initiator, and all are followed by short open reading frames.

Zaret and Sherman (38) have identified a tripartite sequence TAG...TAGT or TATGT..(AT-rich region)..TTT at the 3' end of a set of yeast genes. Based on the study of an iso-1-cytochrome c mutation which deletes this sequence, it appears likely that it is involved in transcription termination, and perhaps polyadenylation. The CDC28 and CDC39 genes have the sequence TAG...TAGT..(AT-rich region)..TTT at positions +53 and +117, respectively. The sequence TATGT...TTT is found at +40, and the sequence TAGT...TTT at +111 of the 3' untranslated region of the CDC36 gene. Both of these sequences lack the first element, TAG, as well as the AT-rich region between the second and third elements. At the 3' end of the CDC37 gene a TAGT...TTT sequence is present at +160. The tripartite sequence lacks the first element, but has an AT-rich region between the second and third elements. The proposed higher eukaryotic polyadenylation sequence AATAAA (39,40) is present in the 3' untranslated regions of the CDC36 (at +26) and CDC37 (at +107) genes. The 2 nucleotides preceding the CDC36 sequence element also conform to the consensus sequence TAAATAAA/G identified by Bennetzen and Hall (41) as occurring at the 3' ends of several sequenced yeast genes, whereas the CDC37 element is preceded by the nucleotides TC. Interesting, the CDC37 gene has at its 3' end a 14 bp sequence which Montgomery et al. (35) have found at the 3' ends of the iso-1 and iso-2-cytochrome c genes. The sequence TTTTTTATTGTTAT at +45 differs by only one nucleotide from the TTTTTTAT/CAGTTAT element reported by Montgomery et al. (35), which occurs at +43 at the 3' end of the iso-2-cytochrome c gene, and at +129 at the 3' end of the iso-1-cytochrome c gene. The significance of the presence of this or a closely related sequence at the 3' end of some yeast genes is unknown.

We thank Dr. Russell Doolittle for performing sequence comparisons. This work was funded by PHS grant GM28005 and NSF grant DCB8402344 to S.I.R. S.I.R. was supported in part by American Cancer Society Faculty Research Award FRA-248.

*Present address: CSIRO-Division of Plant Industry, PO Box 1600, Canberra, ACT 2601, Australia

+Present address: Department of Molecular Biology, Scripps Clinic and Research Foundation, 10666 North Torrey Pines Road, La Jolla, CA 92037, USA

REFERENCES

1. Hartwell, L.H., Mortimer, R.K., Culotti, J., and Culotti, M. (1973). Genetics 74, 267-286.
2. Hartwell, L.H., Culotti, J., Pringle, J.R., and Reid, B.J. (1974). Science 183, 46-51.
3. Reed, S.I. (1980). Genetics 95, 561-577.
4. Pringle, J.R. and Hartwell, L.H. (1982). In Strathern, J.N., Jones, E., and Broach, J.R. (eds), The Molecular Biology of the Yeast Saccharomyces, Cold Spring Harbor Laboratory, New York, pp. 79-142.
5. Hereford, L.M. and Hartwell, L.H. (1974). J. Mol. Biol. 84, 445-461.
6. Hartwell, L.H. (1976). J. Mol. Biol. 104, 803-817.
7. Byers, B. and Goetsch, L. (1975). J. Bacteriol. 124, 511-523.
8. Johnston, G.C., Pringle, J.R., and Hartwell, L.H. (1977). Exp. Cell Res. 105, 79-88.
9. Bucking-Throm, E., Duntze, W., Hartwell, L.H., and Manney, T.R. (1973). Exp. Cell Res. 76, 99-110.
10. Nasmyth, K.A. and Reed, S.I. (1980). Proc. Nat. Acad. Sci. U.S.A. 77, 2119-2123.
11. Breter, H.-J., Ferguson, J., Peterson, T.A., and Reed, S.I. (1983). Mol. Cell. Biol. 3, 881-891.
12. Lorincz, A.T. and Reed, S.I. (1984). Nature 307, 183-185.
13. Reed, S.I., Hadwiger, J.A., and Lorincz, A.T. (1985). Proc. Nat. Acad. Sci. U.S.A. 82, 4055-4059.
14. Peterson, T.A., Yochem, J., Byers, B., Nunn, M.F., Duesberg, P.H., Doolittle, R.F., and Reed, S.I. (1984). Nature 309, 556-558.
15. Chou, P.Y. and Fasman, G.D. (1974). Biochemistry 13, 222-245.
16. Chou, P.Y. and Fasman, G.D. (1978). Adv. Enzymol. 47, 45-147.
17. Kyte, J. and Doolittle, R.F. (1982). J. Mol. Biol. 157, 105-132.
18. Bennetzen, J.L. and Hall, B.D. (1982). J. Biol. Chem. 257, 3026-3031.
19. Sanger, F., Nicklen, S., and Coulsen, A.R. (1977). Proc. Nat. Acad. Sci. U.S.A. 74, 5463-5467.
20. Messing, J., Crea, R., and Seeburg, P.A. (1981). Nucl. Acids Res. 9, 308-321.
21. Henikoff, S. (1984). Gene 28, 351-359.
22. Ish-Horowicz, D. and Burke, J.F. (1981). Nucl. Acids Res. 9, 2989-2998.
23. Rosbash, M., Harris, P.K.W., Woolford, Jr., J.L., and Teem, J.L. (1981). Cell 24, 679-686.
24. Langford, C.J. and Gallwitz, D. (1983). Cell 33, 519-527.
25. Dutcher, S.K. and Hartwell, L.H. (1982). Genetics 100, 175-184.
26. Steinert, P.M., Idler, W.W., and Goldman, R.D. (1980). Proc. Nat. Acad. Sci. U.S.A. 77, 4534-4538.
27. Reed, S.I., de Barros Lopes, M.A., Ferguson, J., Hadwiger, J.A., Ho, J.-Y., Horwitz, R., Jones, C.A., Lorincz, A.T., Mendenhall, M.D., Peterson, T.A., Richardson, S., and Wittenberg, C. (1985). Cold Spring Harbor Symp. Quant. Biol. 50, 627-634.
28. Laughon, A. and Gesteland, R.F. (1984). Mol. Cell. Biol. 4, 260-267.
29. Laughon, A. and Gesteland, R.F. (1982). Proc. Nat. Acad. Sci. U.S.A. 79, 6827-6831.
30. Grosschedl, R. and Birnstiel, M.L. (1980). Proc. Nat. Acad. Sci. U.S.A. 77, 1432-1436.
31. Wasylyk, B., Derbyshire, R., Grey, A., Molko, D., Roget, A., Teoule, R., and Chambon, P. (1980). Proc. Nat. Acad. Sci. U.S.A. 77, 7024-7028.
32. Faye, G., Leung, D.W., Tatchell, K., Hall, B.D., and Smith, M. (1981). Proc. Nat. Acad. Sci. U.S.A. 78, 2258-2262.
33. Sentenac, A. and Hall, B.D. (1982). In Strathern, J.N., Jones, E., and

Broach, J.R. (eds.), The molecular biology of yeast Saccharomyces, Cold Spring Harbor Laboratory, New York, pp. 561-606.
34. Dobson, M.J., Tuite, M.F., Roberts, N.A., Kingsman, A.J., and Kingsman, S.M. (1982). Nucl. Acids Res. 10, 2625-2637.
35. Montgomery, D.L., Leung, D.W., Smith, M., Shalit, P., Faye, G., and Hall, B.D. (1980). Proc. Nat. Acad. Sci. U.S.A. 77, 541-545.
36. Burke, R.L., Tekamp-Olson, P., and Najarian, R. (1983). J. Biol. Chem. 258, 2193-2201.
37. Kozak, M. (1981). Nucl. Acids Res. 9, 5233-5252.
38. Zaret, K.S. and Sherman, F. (1982). Cell 28, 563-573.
39. Proudfoot, N.J. and Brownlee, G.G. (1976). Nature 263, 211-214.
40. Fitzgerald, M. and Shenk, T. (1981). Cell 24, 251-260.
41. Bennetzen, J.L. and Hall, B.D. (1982). J. Biol. Chem. 257, 3018-3025.