Volume 13 Number 23 1985 Nucleic Acids Research

The nucleotide sequence of the DNA ligase gene (*CDC9*) from *Saccharomyces cerevisiae*: a gene which is cell-cycle regulated and induced in response to DNA damage

David G.Barker, Julia H.M.White and Leland H.Johnston

Laboratory of Cell Propagation, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK

Received 27 September 1985; Revised and Accepted 15 November 1985

ABSTRACT

The CDC9 gene of Saccharomyces cerevisiae encodes a DNA ligase, and we have determined the nucleotide sequence of a 3.85 kb fragment of DNA which emcompasses the convergently transcribed CDC9 and CDC36 genes. S1 nuclease mapping has revealed a major 5' end for the CDC9 mRNA, and one major and one minor site for 3' polyadenylation. These two sites lie within the C-terminal coding region of the CDC36 gene, implying that these two genes are transcribed from overlapping sequences. An interesting structural feature of the CDC9 gene is a series of 6 hexanucleotide repeats (ATGATT) which occur within the 650 bp immediately upstream from the site of transcription initiation. These repeat elements may be implicated in the cell division cycle regulated expression of CDC9. Comparison of the predicted amino acid sequence of the yeast DNA ligase (Mr 84,806) with the sequences of the T4 and T7 bacteriophage DNA ligases reveals little similarity except for a stretch of approximately 45 amino acids, comprising 3 short homologous segments. This region may represent an ATP-binding domain common to polynucleotide ligases.

INTRODUCTION

In addition to joining Okazaki fragments during DNA replication, DNA ligase is also required for sealing the nicks in duplex DNA which are generated during recombination and DNA repair processes. The Saccharomyces cerevisiae cell division cycle mutant, cdc9 (1) is defective in all these three functions (2-4), and we have confirmed that the CDC9 gene encodes a DNA ligase both by showing that it can complement a DNA ligase deficient mutant of the fission yeast, Schizosaccharomyces pombe (5), and more recently by demonstrating temperature sensitive DNA ligase activity in several alleles of cdc9 (6). Thus it appears that only one species of DNA ligase is required for DNA replication, repair and recombination in the budding yeast.
From the point of view of regulation, CDC9 would appear to be a most interesting gene. Peterson et al (7) have shown that synthesis of CDC9 mRNA is periodic in the mitotic cell cycle by using cultures of S.cerevisiae which had been synchronised either by size selection or with α-factor. We have

© IRL Press Limited, Oxford, England.

confirmed these results by using feed-starve synchrony (8) and also by elutriation selection (9), and in both cases the CDC9 mRNA level peaks sharply in late G1, just prior to the onset on DNA synthesis (unpublished results). In addition to cell cycle control, Peterson et al (7) have also presented evidence showing that CDC9 mRNA can be induced by irradiating cells with UV light. Our own findings are in broad agreement with this, and additional experiments have shown that CDC9 expression can be induced by a number of other agents which cause damage to cellular DNA (unpublished results).

The S. cerevisiae DNA ligase gene is therefore subject to a complex series of controls, and as the first step towards elucidating the molecular mechanisms underlying these controls we have determined the full nucleotide sequence of CDC9. In addition to identifying non-coding sequences of interest, we have been able to show that the DNA ligase protein shares a homologous domain with other DNA ligases and that this domain may be the site of ATP binding.

MATERIALS AND METHODS

Reagents and Enzymes

[α-^{32}P] TTP for DNA sequencing and S1 probe preparation was obtained from NEN, and specially purified formamide for S1 hybridisations from BDH. Restriction enzymes, Klenow polymerase and S1 nuclease were all supplied by Pharmacia.

DNA Sequencing

DNA to be sequenced was subcloned in both orientations into the filamentous phage vectors, M13mp10 and M13mp11 (10) by means of the polylinker cloning sites. Chain termination sequencing was performed essentially as described by Sanger et al (11) using the 15 nucleotide primer from Biolabs, except that the sequencing reaction was carried out at 29^{o}C. Products of the sequencing reactions were resolved on 6% polyacrylamide/7M urea gels with run times of between 1½-8 h.

RNA preparation and S1 mapping

Total RNA was prepared from the Saccharomyces cerevisiae CDC9^{+} strain CG379 (α, his4-712, leu2-3, leu2-112, trp1-283, ura3-52, ade5) by the method described by Aves et al (12) for Schizosaccharomyces pombe RNA. It was found that the same quantity of cells (~10^{8}) could be treated in 1/10 volume of reagents, so that lysis and extraction could be carried out in 1.5 ml

Eppendorf tubes. The yield of total RNA varied between 100 and 150 μg per 10^8 cells.

Single-stranded DNA probes, uniformly labelled with ^{32}P, were prepared for S1 mapping by the method described by Burke (13). In this procedure, Klenow polymerase is used to extend the normal sequencing primer across a single-stranded M13 template containing the sequence complementary to the desired probe in the presence of [α-^{32}P] labelled nucleotide triphosphate (~300 Ci/mmole). The resulting double stranded DNA is then cleaved with a restriction enzyme distal to the insert, and the radioactive strand isolated by electrophoresis through a 5% polyacrylamide/7M urea gel, followed by elution in 0.5 M NH_4OAc.

For S1 mapping, the ^{32}P-labelled DNA probe (~50,000 cpm) was co-precipitated with total yeast RNA (20-40 μg) and hybridised overnight at 42°C in the presence of 50% formamide. S1 nuclease was added to a final concentration of either 200 or 1000 U/ml as described by Maniatis et al (14), and digestion performed at 30°C for 1 h. After ethanol precipitation, the nucleic acid pellet was dissolved in 10mM Tris HCl/1mMEDTA (pH 7.5), and then electrophoresed on a 6% polyacrylamide sequencing gel alongside a DNA sequencing ladder. Gels were dried and autoradiographed at -70°C for 24-48 h.

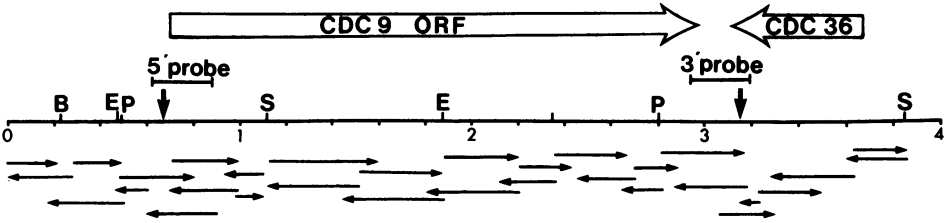

Figure 1. Restriction map and sequence strategy for the 3.8 kb CDC9/CDC36 DNA fragment. The horizontal lines with arrowheads indicate the length and direction of the sequenced M13 clones. The probes used for mapping the 5' end and 3' polyadenylation sites of the CDC9 mRNA are shown, as well as the location of the sites themselves (vertical arrows). The extent of the CDC9 and CDC36 open reading frames are indicated at the top of the figure. Distances along the horizontal line are marked off in 0.2 kb Units, and abbreviations for restriction sites are as follows: B, Bgl II; E, Eco RI; P, Pst I; S, Sst I.

RESULTS

Nucleotide sequence of a DNA fragment containing the CDC9 and CDC36 genes

The S. cerevisiae DNA ligase gene, CDC9, maps to the left arm of chromosome IV within only 1.2 map units of another cell division cycle

```
                20                  40                  60                  80                 100
GATCAAGTAATCTATTACTGTCAATGCCTTTGCGATATGCCTCCAGTATTTGCCCTTGTCGTTAAGTCTTTTGTCCAGCATATCCATGATTTCGAAGAAATCTGCGCTAT

      120                 140                 160                 180                 200
CATATGATTTTTCAGCCAGCTCAATTAACAAGTCCTTGGACACTTGATGATTGTCGTTTGACGTTGCATTTCTTACCAGGACCTGGGTAGAAGAGTACCCTTTCACCAAA

      220                 240                 260                 280                 300                 320
TTCTTCGCAGATCTAACAAATTGTTTCGACATGATTTTTCTGCCTCTTGCCCCCTTTCCTTCTGATCCTGGCACTCATACGAACGGCTCTTTCTTCTACTTCTTAAAACT

      320                 340                 360                 380                 400                 420
CCTATTTGTCATTGAAAGAAACTTTCATCATTACCCGGATGATTCTCTATACGTAGCCACAATTTCAGACAACCTACTTCCAAAATCAAGTACTTTGAACAACATAAGGC

      420                 440                 460                 480                 500                 520
CCCACCAACCTTCCGTAAGACTTGCAATAGGAATTCGTACTGCAGCCGCTGGTTCTATTGCTGTTTAACTCTGTCATCTTTTTGCTTGCTTCGATGTGTTCCTCCACTTG

      520                 540                 560                 580                 600                 620
AGATGGCTGATGGAATTTTTCACCTTAACGCCAAAACCCGTGAAAGTGAAATTTAGCTGCAATAAGCCATGCGTTTGCATCTCGCTGTACTTTGTTCAAAAATAAAAGTA

      620                 640                 660                 680                 700                 720
ATGATTTTGAAGGTTTAGAGCAGCTTTTAACGTCATTGTTCATCAATTACATGCGCAGATTACTGACCGGTTGCCTTCTTTCATCTGCACGTCCCTTGAAATCAAGATTG
                                                  M   R   R   L   L   T   G   C   L   L   S   S   A   R   P   L   K   S   R   L
                                                                  CDC 9
      720                 740                 760                 780                 800                 820
CCATTATTGATGTCATCGTCATTACCTTCCTCTGCCGGTAAGAAGCCTAAACAAGCCACTTTGGCTAGATTCTTCACTTCCATGAAAAATAAGCCAACAGAAGGTACTCC
  P   L   L   M   S   S   S   L   P   S   S   A   G   K   K   P   K   Q   A   T   L   A   R   F   F   T   S   M   K   N   K   P   T   E   G   T   P

      820                 840                 860                 880                 900                 920
TTCTCCCAAAAAATCATCCAAACATATGCTGGAAGAGAGAATGGATAATGTTAGTGGGGAAGAGGAATACCGCGACCAAGAAATTGAAACAGACGGCTGTTACACATACTG
  S   P   K   K   S   S   K   H   M   L   E   E   R   M   D   N   V   S   G   E   E   E   Y   A   T   K   K   L   K   Q   T   A   V   T   H   T

      920                 940                 960                 980                1000                1020
TAGCAGCTCCAAGCTCCATGGGTAGCAATTTTTCTTCTATACCATCATCCGCTCCCTCTTCCGGTGTTGCTGATTCACCACAACAATCTCAGAGGTTGGTAGGTGAAGTT
  V   A   A   P   S   S   M   G   S   N   F   S   S   I   P   S   S   A   P   S   S   G   V   A   D   S   P   Q   Q   S   Q   R   L   V   G   E   V

     1020                1040                1060                1080                1100                1120
GAAGACGCCTTGAGCTCAAATAATAATGATCACTATTCGTCTAATATCCCCTATTCTGAAGTTTGTGAGGTTTTTAACAAGATTGAGGCCATATCTTCCCGTTTAGAGAT
  E   D   A   L   S   S   N   N   N   D   H   Y   S   S   N   I   P   Y   S   E   V   C   E   V   F   N   K   I   E   A   I   S   S   R   L   E   I

     1120                1140                1160                1180                1200                1220
TATAAGAATCTGTTCTGATTTCTTTATTAAGATAATGAAGCAATCGTCTAAGAACGTAATACCTACAACATACCTTTTCATCAATAGATTGGGCCCTGACTACGAAGCAG
  I   R   I   C   S   D   F   F   I   K   I   M   K   Q   S   S   K   N   V   I   P   T   T   Y   L   F   I   N   R   L   G   P   D   Y   E   A

     1220                1240                1260                1280                1300                1320
GCTTGGAATTGGGCCTTGGAGAAAATCTTCTTATGAAAACCATAAGCGAAACTTGCGGGAAATCCATGAGCCAAATTAAACTTAAATACAAGGATATTGGGGATTTGGGT
  G   L   E   L   G   L   G   E   N   L   L   M   K   T   I   S   E   T   C   G   K   S   M   S   Q   I   K   L   K   Y   K   D   I   G   D   L   G

     1320                1340                1360                1380                1400                1420
GAAATAGCAATGGGTGCGAGAAATGTCCAACCCACTATGTTTAAGCCCAAACCCTTGACCGTGGGTGAGGTCTTCAAAAATCTTAGAGCTATTGCTAAGACTCAGGGCAA
  E   I   A   M   G   A   R   N   V   Q   P   T   M   F   K   P   K   P   L   T   V   G   E   V   F   K   N   L   R   A   I   A   K   T   Q   G   K

     1420                1440                1460                1480                1500                1520
AGATTCACAATTAAAAAAGATGAAGCTGATCAAGAGAATGCTTACTGCTTGCAAAGGTATAGAGGCTAAGTTTTTGATTAGATCCCTAGAGTCAAAATTGAGAATCGGTC
  D   S   Q   L   K   K   M   K   L   I   K   R   M   L   T   A   C   K   G   I   E   A   K   F   L   I   R   S   L   E   S   K   L   R   I   G

     1520                1540                1560                1580                1600                1620
TTGCTGAAAAGACTGTCTTAATATCCCTATCGAAGGCTCTTTTGCTCCATGATGAAAATAGGGAGGACTCTCCAGACAAAGATGTCCCCATGGATGTGTTAGAAAGTGCT
  L   A   E   K   T   V   L   I   S   L   S   K   A   L   L   L   H   D   E   N   R   E   D   S   P   D   K   D   V   P   M   D   V   L   E   S   A

     1620                1640                1660                1680                1700                1720
CAACAAAAGATAAGAGACGCATTTTGTCAAGTACCCAACTATGAAATTGTTATAAACTCATGCCTGGAACACGGTATTATGAATTTAGATAAATATTGTACTCTAAGACC
  Q   Q   K   I   R   D   A   F   C   Q   V   P   N   Y   E   I   V   I   N   S   C   L   E   H   G   I   M   N   L   D   K   Y   C   T   L   R   P

     1720                1740                1760                1780                1800                1820
AGGAATTCCATTGAAACCCATGTTAGCCAAGCCTACCAAGGCAATTAATGAAGTATTGGATAGATTTCAAGGAGAAACTTTTACGTCAGAATACAAATACGATGGTGAAA
  G   I   P   L   K   P   M   L   A   K   P   T   K   A   I   N   E   V   L   D   R   F   Q   G   E   T   F   T   S   E   Y   K   Y   D   G   E

     1820                1840                1860                1880                1900                1920
GGGCTCAGGTGCATTTACTAAATGATGGTACAATGAGGATTTATTCCAGAAATGGCGAGAATATGACTGAGAGATATCCAGAAATTAACATAACGGATTTTATCCAGGAT
  R   A   Q   V   H   L   L   N   D   G   T   M   R   I   Y   S   R   N   G   E   N   M   T   E   R   Y   P   E   I   N   I   T   D   F   I   Q   D
```

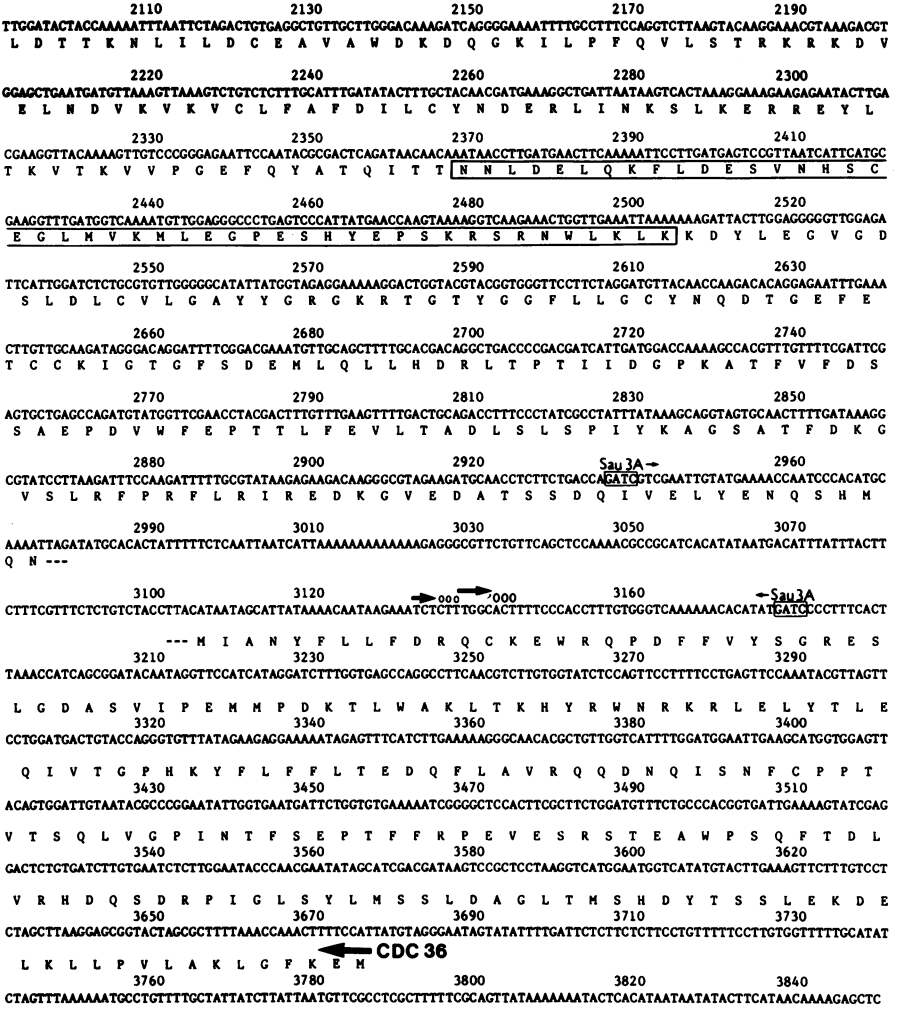

Figure 2. Nucleotide sequence of the CDC9/CDC36 gene region. For reasons of clarity, only the coding strand of the CDC9 gene is shown, with translation starting at the first in-frame ATG codon (res.711-713). Alternative ATG initiation codons are indicated by dashed underlining. The translated amino acid sequence of the convergently transcribed CDC36 gene is also shown, starting at residue 3674 and continuing backwards through to residue 3104. The Rsa I (res. 637-874) and Sau 3A (res. 2936-3177) fragments used in S1 mapping experiments are indicated on the sequence, as well as the positions of the 5' end and 3' polyadenylation sites of the CDC9 mRNA (open circles). Filled circles mark the six hexanucleotide repeats ATGATT within the CDC9 upstream sequence, and overlined sequences indicate the region of dyad symmetry. The 46 amino acids which are boxed are those which show homology to regions of the T4 and T7 bacteriophage DNA ligases, and include the putative ATP binding site (Fig. 4).

gene, the 'start' mutant cdc36 (15). Peterson et al (7) have found that a 4.6 kb fragment of yeast DNA is capable of complementing both cdc9 and cdc36 mutations, and they showed by subcloning analysis and strand specific hybridisation that the two genes lie immediately adjacent to each other on the chromosome and are transcribed convergently off opposite DNA strands. R-loop mapping was used to estimate the lengths of the CDC36 and CDC9 transcripts as 615 and 2,500 bp respectively(16).

We have independently isolated a 6.3kb fragment of S.cerevisiae DNA which is able to fully complement the cdc9 mutation (5), and whose restriction site pattern matches that of the 4.6kb fragment of Peterson et al(7). Sequencing of a 3.8kb region which spans the CDC9/CDC36 locus was carried out using the M13/dideoxy chain termination method, generating the series of overlapping sequences shown in Fig. 1. Greater than 90% of the entire sequence is covered on both strands, and there is no ambiguity within those short stretches of sequence which have only been determined on a single strand. Analysis of the sequence reveals two open reading frames (ORFs) of 641 bp and 2375 bp in length, located on opposing strands of the DNA, and separated by only 121 bp (Fig. 2.). The correlation with both the sizes of the CDC9 and CDC36 transcribed regions given above and the relative orientation of their transcription argues overwhelmingly that these two ORFs correspond to the structural components of the two genes. Furthermore, the amino acid sequence of the short ORF exactly matches the structure of the CDC36 gene product as deduced from the sequence of an independently cloned fragment of DNA (17).

The coding region of the CDC9 gene

The larger open reading frame, corresponding to the CDC9 gene, extends from residue 600 to residue 2975, and contains no less than 5 possible ATG initiation codons between residues 711-921 (Fig. 2). Transcription mapping (see below) places the major 5' end of the mRNA transcript at nucleotide position 671-673, and since it is usually the first ATG codon downstream from the transcription start which is recognised by the translation machinery (18), initiation of translation probably occurs at the first ATG in the ORF (res. 711-713). Although the S.cerevisiae DNA ligase has never been purified to homogeneity, we have been able to obtain a direct estimate of the size of the monomer polypeptide by means of the covalent AM^{32}P-DNA ligase adduct, which can be readily prepared in crude yeast cell extracts (19). Analysis by SDS polyacrylamide gel electrophoresis gives a molecular weight of 88,000 daltons for the S. cerevisiae ligase-AMP adduct, which agrees moderately well with a

Table 1.

Amino Acid	Codon	CDC9 No. Used	CDC9 %Codon Usage	CDC36 No. Used	CDC36 %Codon Usage	ADH1 %Codon Usage	Amino Acid	Codon	CDC9 No. Used	CDC9 %Codon Usage	CDC36 No. Used	CDC36 %Codon Usage	ADH1 %Codon Usage
Phe(F)	TTT	18	62%	8	57%	0%	Tyr(Y)	TAT	11	48%	6	86%	0%
Phe	TTC	11	38	6	42	100	Tyr	TAC	12	52	1	14	100
Leu(L)	TTA	14	17	6	26	8	Term	TAA	-		-		-
Leu	TTG	32	39	4	17	79	Term	TAG	-		-		-
Leu	CTT	18	22	3	13	0	His(H)	CAT	6	60	1	25	9
Leu	CTC	3	4	5	22	0	His	CAC	4	40	3	75	91
Leu	CTA	8	10	5	22	13	Gln(Q)	CAA	16	62	8	80	100
Leu	CTG	7	9	0	0	0	Gln	CAG	10	38	2	20	0
Ile(I)	ATT	18	43	4	67	42	Asn(N)	AAT	23	70	4	80	0
Ile	ATC	8	19	2	33	57	Asn	AAC	10	30	1	20	100
Ile	ATA	16	38	0	0	0	Lys(K)	AAA	40	62	5	50	17
Met(M)	ATG	22	100	6	100	100	Lys	AAG	25	38	5	50	83
Val(V)	GTT	16	42	1	11	53	Asp(D)	GAT	31	69	7	58	13
Val	GTC	9	24	2	22	47	Asp	GAC	14	31	5	42	87
Val	GTA	8	21	4	44	0	Glu(E)	GAA	38	68	11	77	100
Val	GTG	5	13	2	22	0	Glu	GAG	18	32	2	23	0
Ser(S)	TCT	15	23	0	0	67	Cys(C)	TGT	8	50	0	0	100
Ser	TCC	17	26	5	31	33	Cys	TGC	8	50	2	100	0
Ser	TCA	16	25	5	31	0	Term	TGA	-		-		-
Ser	TCG	6	9	3	19	0	Trp(W)	TGG	3	100	4	100	100
Pro(P)	CCT	10	29	3	25	15	Arg(R)	CGT	4	11	1	9	0
Pro	CCC	10	29	1	8	8	Arg	CGC	1	3	0	0	0
Pro	CCA	13	37	5	42	77	Arg	CGA	0	0	1	9	0
Pro	CCG	2	6	3	25	0	Arg	CGG	0	0	0	0	0
Thr(T)	ACT	19	41	4	31	36	Ser	AGT	6	9	2	13	0
Thr	ACC	8	17	3	23	64	Ser	AGC	5	8	1	6	0
Thr	ACA	11	24	4	31	0	Arg	AGA	23	62	5	45	100
Thr	ACG	8	17	2	15	0	Arg	AGG	9	24	4	36	0
Ala(A)	GCT	16	43	2	29	54	Gly(G)	GGT	19	42	4	57	93
Ala	GCC	6	16	2	29	46	Gly	GGC	8	18	1	14	7
Ala	GCA	12	32	1	14	0	Gly	GGA	10	22	2	29	0
Ala	GCG	3	8	2	29	0	Gly	GGG	8	18	0	0	0

Codon usage of the CDC9 and CDC36 genes. The percentage codon usage of the two genes can be compared with the highly expressed ADH1 gene (19). The single letter amino acid coding used in Figs. 1 and 4 is shown alongside the conventional three letter coding.

calculated molecular weight of 84,806 daltons based on the 755 amino acids between the first of the ATG codons (res.711-713) and the end of the open reading frame.

Bennetzen and Hall (20) have shown that there is a correlation between the extent of codon bias within a gene and the level of both mRNA and protein expression. An examination of the codon usage of the CDC9 gene reveals very little bias towards preferred codons (Table 1). Using the criteria of Bennetzen and Hall (20), the CDC9 message should only be present at a level of approximately 0.01% of total polyadenylated RNA, and indeed this agrees well with our own estimates based on the S1 mapping experiments described in the next section. The codon usage of the CDC36 gene product is also without

bias(Table 1), and this is consistent with the very low level of CDC36 transcription (16).

Transcription mapping of the CDC9 gene

Total RNA was prepared from the S.cerevisiae strain CG379 and used in S1 mapping experiments to determine the 5' end(s) and 3' polyadenylation site(s) of the mature CDC9 message. The approximate location of these sites was first established by the traditional Berk and Sharp procedure (21), using agarose gels and Southern hybridisation to measure the size of the protected fragments (results not shown). The various DNA fragments used in these experiments together spanned the entire CDC9 gene, and the analysis of the S1 digested hybrids on alkaline agarose gels demonstrated that CDC9 mRNA is un-spliced. This is consistent with the absence of the S.cerevisiae intron consensus sequence TACTAAC (22) within the coding region. Precise mapping of the CDC9 transcript was then carried out in the following way :

(i) Mapping the 5' end. An Rsa I fragment of 237 bp in length and spanning the 5' end of the mRNA (Fig.2) was cloned into the Sma I site of M13mp11 in one orientation, and the complementary strand was then uniformly labelled with ^{32}P using universal primer and Klenow polymerase (see Materials and Methods). The single strand radiolabelled probe was purified away from complementary sequence by restriction digestion and polyacrylamide/urea gel electrophoresis and hybridised with total S.cerevisiae RNA. Two different concentrations of S1 were used to digest the DNA/RNA hybrid, which was then denatured and electrophoresed in a 6% polyacrylamide/7M urea gel (Fig. 3). This experiment shows that CDC9 mRNA has only one major 5' end (res. 671-673), and we suspect that the faster migrating and fainter series of bands represent S1 digestion within those regions of the hybrid with high AT composition (e.g.res.685-690).

(ii) Mapping the 3' end. The 3' polyadenylation site of the CDC9 mRNA was mapped in a similar manner using a radioactively labelled DNA probe prepared from a 241bp Sau3A cloned fragment (Fig.2). The S1 digestion patterns are shown in Fig. 3, from which it appears that there are major and minor protected DNA fragments of 207-209 and 200-202 bp in size respectively. Higher S1 concentrations do not convert the longer fragment to the shorter one, suggesting that there are indeed two distinct polyadenylation sites. Surprisingly, both of these sites fall within the coding region of the CDC36 gene (res. 3135-3137 and 3142-3144), thus

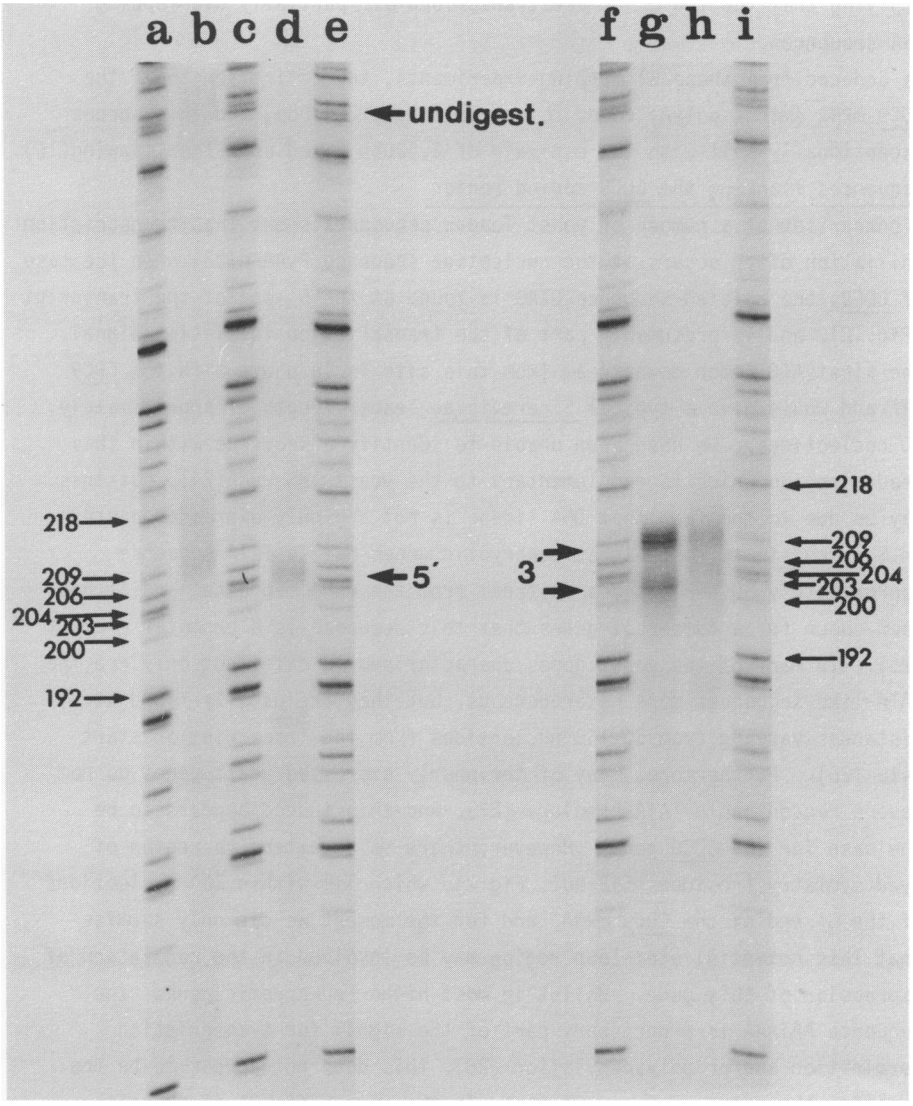

Figure 3. S1 mapping the 5' end and 3' polyadenylation sites of CDC9 mRNA. The 5' terminus of the CDC9 transcript was mapped by the S1 nuclease method using a single-strand probe prepared from the 237 bp Rsa I fragment (see text and Fig. 2). Tracks b and d show the results of digesting the RNA/DNA hybrid with 200 and 1000 U/ml S1 nuclease respectively. Undigested full length probe is shown at the top of the gel in track b. The 3' polyadenylation sites were located using the 241 bp Sau 3A fragment probe (Fig. 2), with tracks g and h representing S1 nuclease digestion at 200 and 1000 U/ml respectively. Molecular weight marker bands shown in tracks a, c,e,f and i are derived from a single dideoxy sequencing reaction.

implying that the two mRNAs are transcribed off partially overlapping DNA sequences.

As deduced from these S1 mapping experiments, the entire length of the CDC9 mRNA (minus polyA) comes to approximately 2,470bp, and this agrees exceptionally well with the estimate of 2,500bp based on R-loop mapping(16).

Sequences flanking the CDC9 coding region

A comparison of a number of yeast leader sequences shows that transcription initiation often occurs at the nucleotide sequence PyAAPu(23). In the case of CDC9, the related sequence GAAG is found at the 5' end of the transcript (Fig. 2), and is presumably part of the transcription initiation signal. The first ATG codon downstream from this site is in phase with the CDC9 ORF and would give a typical S.cerevisiae leader length of approximately 40 nucleotides. We have been unable to identify a sequence within this leader region which is complementary to the yeast 18S rRNA(24), but this may be due to the fact that DNA ligase is not a highly expressed protein in S.cerevisiae. Most higher eukaryotic genes have a TATA sequence approximately 30 nucleotides upstream from the mRNA cap site, and it has been shown for a number of genes that this sequence is a promoter element (25). In many of the yeast genes characterised to date, not only are the TATA-like sequences more heterogeneous, but they are usually found at distances varying from 50-150 nucleotides from the transcription start site (26). Furthermore, many of the poorly expressed yeast genes do not have a recognisable TATA homology (27), and this indeed appears to be the case for the CDC9 gene. However, there is an extensive region of dyad symmetry (residues 568-601, Fig.2), which lie within 100 nucleotides of the 5' end of the CDC9 mRNA, and for the moment we can only surmise that this potential stem-loop region may be involved in the regulation of expression of this gene. Whilst in most higher eukaryotic genes the sequence AATAAA is a necessary part of the signal for transcription termination and/or polyadenylation (28), this does not appear to be the case for the genes of S.cerevisiae. In the search for 3' flanking sequences which are common to yeast genes, Bennetzen and Hall (29) have identified the consensus TAAATAA$^{A}_{G}$, which occurs 28-33bp before the polyadenylation site of several genes. Although the 3' flanking region of the CDC9 gene does contain the related sequence AATAAG (Fig. 2) it is too close to the two polyadenylation sites (13 and 20bp) to fit the above consensus. Similarly, the consensus suggested by Zaret and Sherman (30), comprising the series of short elements of the form TAG...$^{TAGT}_{TATGT}$....TTT is also

not present within the 3' flanking sequence of the CDC9 transcript. Although it is not unprecedented for yeast genes to lack either consensus, the fact that transcription termination of the CDC9 gene takes place within the CDC36 coding region may impose additional constraints upon the sequence of a termination signal.

DISCUSSION

We have determined the sequence of the CDC9 gene from budding yeast, and analysis of the structure of this gene suggests that it encodes an enzyme of 755 amino acids in length and with a polypeptide molecular weight of 84,806 daltons. In those organisms (not including yeast) where DNA ligase has been purified to homogeneity, the enzyme behaves as a monomeric species, varying in molecular weight between 41 and 79 kD in prokaryotes and bacteriophage (31,32) and between 90 and 200 kD in higher eukaryotes (33). The Saccharomyces cerevisiae DNA ligase does not really fall into either size category, but the existence of only a single species in yeast (see Introduction) probably places it closer in functional terms to the prokaryotic enzymes, since in higher eukaryotes there is evidence that two species of DNA ligase are required to carry out the multiple roles of the enzyme (33).

As regards sequence comparison, only those of the bacteriophage T7 and T4 DNA ligases (M_r 41,103 and 55,230 respectively) are currently available (31,34). Computer assisted searches have failed to reveal extensive homologies between the yeast and the two bacteriophage sequences, and indeed Armstrong et al (34) have only succeeded in aligning 80 amino acid residues between the T4 and T7 sequences by introducing numerous gaps into both sequences. More interestingly, searches for limited local homologies between the yeast and 'phage enzymes have identified a stretch of approximately 45 amino acids with the characteristics of a conserved structural domain (Fig. 4). By introducing only two small gaps, residues 326-367 of the T4 DNA ligase sequence can be aligned with residues 552-597 of the yeast enzyme to give a region with an overall homology of greater than 50% (including conservative amino acid differences). This region can be divided into three highly homologous segments, of which the most striking is the central block with 9 out of 10 matching amino acids. Part of the T7 DNA ligase sequence can also be aligned with this region (res.198-239), and this has the effect of further emphasising the three conserved segments. It is surely not coincidental that this region falls within the stretch of closest homology between the two bacteriophage enzymes

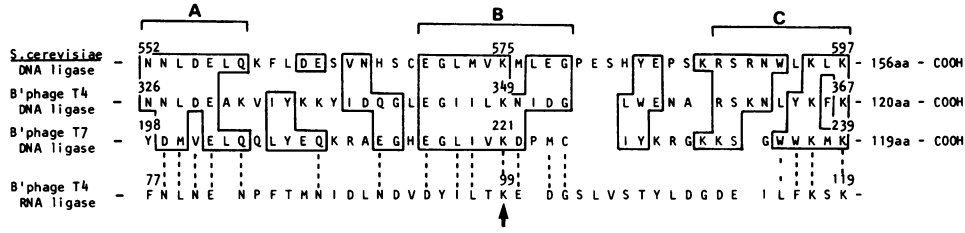

Figure 4. Amino acid homologies between three DNA ligases and a possible alignment with a region of the bacteriophage T4 RNA ligase. 46 amino acids of the S.cerevisiae DNA ligase which show homology within three closely spaced regions (A,B,C) with the bacteriophage T4 and T7 DNA ligases. The sequence which flanks the ATP binding residue Lys^{99} (arrowed) of the bacteriophage T4 RNA ligase can be partially aligned with these three homologous regions.

(34), and that it is located at a similar distance from the C-terminus of all three DNA ligases.

Could this homologous region correspond to a catalytic domain? A clue to this question may have come from some recent work by Thogersen et al (35), who have identified the lysine residue (Lys^{99}) in T4 RNA ligase which binds AMP to form the covalent ligase-AMP adduct. DNA ligases also form covalent AMP adducts via the ε-amino group of a lysine residue, and it is thought that this is a genuine intermediate in the overall ligation reaction (36). Comparing the amino acid sequence which flanks the AMP-binding lysine residue in T4 RNA ligase with the region of homology between the DNA ligases (Fig. 4), it is possible to obtain limited amino acid matches with all three conserved segments when Lys^{99} is aligned with Lys^{575}, Lys^{349}, Lys^{221} of the S.cerevisiae, T4 and T7 DNA ligases respectively. It is therefore possible that this whole homologous region forms part of the ATP binding domain of RNA and DNA ligases, and that these lysine residues are the sites of covalent nucleotide binding. If this turns out to be the case, then this will represent a new class of nucleotide binding domains, in addition to those which have already been identified for protein kinases (37), for dehydrogenases (38) and for other adenine nucleotide binding proteins (39,40). Given the ease with which T4 DNA ligase can be purified (41) it should prove relatively straightforward to identify the AMP-binding residue, and hence test this hypothesis.

Although the genes on yeast chromosomes are in general closely packed together, as far as we are aware, this is the first reported example of overlapping convergent transcription from two adjacent genes. Both of the 3' polyadenylation sites of CDC9 lie within the C-terminal coding region of CDC36, and although

we have not mapped the position of 3' polyadenylation for the CDC36 gene, transcription probably overlaps by at least 50-60 nucleotides taking into account the 3' untranslated region of the CDC36 transcript. Convergent overlapping transcription has been studied in bacteria by Ward and Murray (42) using λtrp phage constructs. They have shown that when the promoters are strong the level of either transcript can be significantly reduced, and the authors propose that colliding RNA polymerase molecules may be responsible for this effect. Even if the situation were similar in yeast, the rates of transcription from either CDC36 or the transiently expressed CDC9 (see below) are unlikely to be sufficient to interfere with the expression of either gene. CDC36 is included within the category of so-called 'start' genes, encoding proteins which are required in early G1, when the cell is faced with the option of initiating a further cell cycle (43). However, unlike CDC9, CDC36 transcription is maintained at a constant level throughout the cell cycle (7), and from a functional point of view it is difficult to imagine that the overlapping transcription of these two genes can be other than fortuitous. CDC9 is one of the very few yeast genes so far discovered whose transcription is regulated by progress through the mitotic cycle (7). The others include the histone H2A and H2B genes (44), the TMP1 gene which encodes thymidylate synthase (45) and the HO gene which codes for an endonuclease required for mating type switching (46). At present, little is known about the mechanism governing cell cycle-dependent gene expression but Nasmyth (47) has reported that a repeated heptanucleotide sequence in the regulatory region of the HO gene appears to be involved in its specific expression during late G1 of the cell cycle. A short repeat element has also been implicated in the expression of the H2A/H2B genes, although this element has a different sequence from that of the HO gene and is present in only three copies (48). It is therefore particularly interesting that we should find a hexanucleotide sequence ATGATT occuring 6 times within the 650 bp of sequence immediately upstream from the CDC9 5' leader region (Fig. 2). Again, the sequence of this element differs from either the HO or histone gene repeats, but this could indicate that whilst broadly similar mechanisms might exist for regulating cell cycle specific gene transcription, there may be considerable variation in the fine details of these controls. In fact, we have evidence that there are significant differences in the exact timings of CDC9 and H2A/H2B expression within the S.cerevisiae mitotic cell cycle (unpublished results). On the other hand, we have also found that the TMPI mRNA level fluctuates in precisely the same part of the cell cycle as that of CDC9, and it will be interesting to see

whether the same ATGATT sequence occurs in both of these genes.
In addition to cell cycle-dependent regulation, transcription of the CDC9 gene also responds to signals resulting from damage to cellular DNA (7). Preliminary experiments have shown that the 650 bp of DNA upstream from the transcribed region of the CDC9 gene are sufficient for the induction of β-galactosidase activity by UV light and other DNA damaging agents in translational gene fusions between CDC9 and the E.coli lac Z gene (unpublished results). These fusions should prove useful in studying the cell cycle and DNA damage responses of the CDC9 gene, and in showing whether these two forms of regulation are independent.

ACKNOWLEDGEMENTS

We would like to thank Paul Russell for advice on the preparation of single-stranded DNA probes. Our appreciation also to Don Williamson for a critical reading of the manuscript, and to Mary Ann Osley for communicating results prior to publication.

REFERENCES

1. Hartwell, L.H., Mortimer,R.K., Culotti,J. and Culotti,M. (1973) Genetics 74, 267-286.
2. Johnston,L.H., and Nasmyth,K.A. (1978) Nature 274, 891-893.
3. Fabre,F. and Roman,H. (1979) Proc. Natl. Acad. Sci. USA. 76, 4586-4588.
4. Johnston,L.H. (1979) Mol. Gen. Genet. 170, 89-92.
5. Barker,D.G. and Johnston,L.H. (1983) Eur. J. Biochem. 134, 315-319.
6. Barker,D.G., Johnson,A.L. and Johnston,L.H. (1985) Mol. Gen. Genet. 200, 458-462.
7. Peterson,T.A., Prakash,L., Prakash,S., Osley,M.A. and Reed,S.I. (1985) Mol. Cell. Biol. 5, 226-235.
8. Williamson,D.H. and Scopes,A.W. (1962) Nature 193, 256-257.
9. Creanor,J. and Mitchison,J.M. (1979) J. Gen. Microbiol. 112, 385-388.
10. Norrander,J., Kempe,T. and Messing,J. (1983) Gene 26, 101-106.
11. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) Proc. Natl. Acad. Sci. USA. 74, 5463-5467.
12. Aves,S.J., Durkacz,B.W., Carr,A. and Nurse,P. (1985) EMBO J. 4, 457-463.
13. Burke, J.F. (1984) Gene 30, 63-68.
14. Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) In Molecular cloning: a laboratory manual, pp. 207-209, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.
15. Shuster,J.R. (1982) Mol. Cell. Biol. 2, 1052-1063.
16. Breter,H.J., Ferguson,J., Peterson,T.A. and Reed,S.I. (1983) Mol. Cell. Biol. 3, 881-891.
17. Peterson,T.A., Yochem,J., Byers,B., Nunn,M.F., Duesberg,P.H., Doolittle R.F. and Reed,S.I. (1984) Nature, 556-558.
18. Kozak,M. (1978) Cell 15, 1109-1123.
19. Banks,G.R. and Barker,D.G. (1985) Biochim. Biophys. Acta (in press).
20. Bennetzen,J.L. and Hall,B.D. (1982) J. Biol. Chem. 257, 3026-3031.
21. Berk,A.J. and Sharp,P.A. (1977) Cell 12, 721-732.

22. Langford,C.J. and Gallwitz,D. (1983) Cell 33, 519-527.
23. Burke,R.L., Tekamp-Olsen,P. and Najarian,R. (1983) J. Biol. Chem. 258, 2193-2201.
24. Shine,J. and Dalgarno.L. (1974) Biochem. J. 141, 609-615.
25. Benoist,C., O'Hare,K., Breathnach,R. and Chambon,P. (1980) Nuc. Acids Res. 8, 127-142.
26. Reynolds,P., Higgins,D.R., Prakash,L. and Prakash,S. (1985) Nuc. Acids Res. 13, 2357-2372.
27. Laughon,A. and Gesteland,R.F. (1984) Mol. Cell. Biol. 4, 260-267.
28. Proudfoot,N.J. and Brownlee,G.G. (1976) Nature 263, 211-214.
29. Bennetzen,J.L. and Hall, B.D. (1982) J. Biol. Chem. 257, 3018-3025
30. Zaret, K.S. and Sherman,F. (1982) Cell 28, 563-573.
31. Dunn,J.J. and Studier,F.W. (1981) J. Mol. Biol. 148, 303-330.
32. Takahashi,M., Yamaguchi,E. and Uchida,T. (1984) J. Biol Chem 259, 10041-10047.
33. Söderhäll,S. and Lindahl,T. (1976). FEBS Lett. 67, 1-7.
34. Armstrong,J., Brown,R.S. and Tsugita,A. (1983) Nuc. Acids Res. 11, 7145-7156.
35. Thogerson,H.C., Morris,H.R., Rand,K.N. and Gait,M.J. (1985) Euro. J. Biochem. 147, 325-329.
36. Lehman,I.R. (1974) Science 186, 790-797.
37. Barker,W.C. and Dayhoff,M.O. (1982) Proc. Natl. Acad. Sci. USA 79, 2836-2839.
38. Buener,M., Ford,G.C., Moras,D., Olsen,K.W. and Rossman,M.G. (1974) J. Mol. Biol. 90, 25-49.
39. Walker,J.E., Saraste,M., Runswick,M.J. and Gay,N.J. (1982) EMBO J. 1, 945-951.
40. Wierenga,R.K. and Hol,W.G.J. (1983) Nature 302, 842-844.
41. Murray,N.E., Bruce,S.A. and Murray K. (1979) J. Mol. Biol. 132, 493-505.
42. Ward,D.F. and Murray,N.E. (1979) J. Mol. Biol. 133, 249-266.
43. Reed,S.I. (1980) Genetics 95, 561-577.
44. Hereford,L.M. and Osley,M.A. (1981) Cell 24, 367-375.
45. Storms,R.K., Ord,R.W., Greenwood,M.T., Mirdamadi,B., Chu,F.K. and Belfort,M. (1984) Mol. Cell. Biol. 4, 2858-2864.
46. Nasmyth,K. (1983) Nature 302, 670-676.
47. Nasmyth,K. (1985) Cell 42, 225-235.
48. Osley,M.A., Gould,J., Kim,S., Kane,M. and Hereford,L. Cell (submitted)