Volume 12 Number 22 1984 Nucleic Acids Research

A comparison of yeast ribosomal protein gene DNA sequences

John L.Teem[1], Nadja Abovich[1], Norbert F.Kaufer[2], William F.Schwindinger[2], Jonathan R.Warner[2], Allison Levy[3], John Woolford[3], R.J.Leer[4], M.M.C.van Raamsdonk-Duin[4], W.H.Mager[4], R.J.Planta[4], L.Schultz[5], J.D.Friesen[5], Howard Fried[6] and Michael Rosbash[1]

[1]Department of Biology, Brandeis University, Waltham, MA 02254, USA

Received 16 July 1984; Revised and Accepted 9 October 1984

ABSTRACT

The DNA sequences of eight yeast ribosomal protein genes have been compared for the purpose of identifying homologous regions which may be involved in the coordinate regulation of ribosomal protein synthesis. A 12 bp homology was identified in the 5' DNA sequence preceding the stuctural gene for 6 out of 8 yeast ribosomal protein genes. In each case the homologous sequence was found at a position approximately 300 bp preceding the transcription start of the ribosomal protein gene. This homology was not identified in any non-ribosomal protein gene examined. Additional homologies between ribosomal protein genes were identified in the transcribed regions, including the untranslated 5' and 3' DNA regions flanking the coding regions.

INTRODUCTION

The synthesis of ribosomal proteins occurs coordinately in yeast (1,2) as well as in other eukaryotes (3). The basis for this coordinate regulation is unknown, although it is likely to be mediated at both the transcriptional (4) and post-transcriptional level (5). Presumably yeast ribosomal protein genes share some common features which allow the expression of the approximately 75 ribosomal proteins to be coordinately regulated. An analysis of the ribosomal protein gene structure at the DNA sequence level would provide an indication of the extent to which these genes are similar, and perhaps reveal putative regulatory elements. To this end, eight yeast ribosomal protein genes have been compared for the purpose of identifying sequences that may be specific to ribosomal protein genes as a group.

METHODS

Sequence Comparison Strategy

The eight ribosomal protein genes compared include six genes isolated from *S. cerevisiae*, and two from *S.*

carlsbergensis, and are listed in Table 1. Non-ribosomal protein genes which were analyzed are also listed in Table 1. Several procedures have been adopted to simplify the comparisons. For example, all the DNA sequences upstream from the start methionines were placed end to end within a single DNA sequence data file. This composite sequence was then compared to itself in order to obtain a single output (instead of a collection of outputs from many pairwise combinations of individual 5' ends). Composite sequences of the other regions were made and compared in the same way. The composite files define (1) 5' flanking regions (sequences upstream from the start methionines), (2) 3' flanking regions (sequences downstream from the termination codons), (3) coding regions and (4) introns (when applicable). To further simplify the analysis, the two S. carlsbergensis ribosomal protein gene DNA sequences were not included in the initial homology search so as not to exclude homologies that might be specific to the S. cerevisiae genes.

Each composite sequence was compared to itself using a forward homology matrix program (15). This program was used to identify regions of the composite sequence having homology of 8 (or more) bases matching within a span of an 11 base region. The parameters of the program were empirically set (Range=5, Scale=.95, Minimum value plotted=75) such that the 8 matching bases must have at least four consecutive matches within the eleven base interval, and no consecutive mismatching bases. If a homology meeting these criteria was identified in four of the six S. cerevisiae ribosomal protein genes, then the region of homology was used as a subsequence in a second comparison to search the remaining ribosomal protein genes for weaker homologies that might have been missed in the first search. (The second comparison also required that additional matches have at least 75% homology to the subsequence.) Homologous sequences were then compared to form a consensus sequence, and non-ribosomal protein yeast genes (Table 1) were then searched for each consensus.

The 5' end of RP51A was compared with the 5' end of RP51B using the forward homology matrix program (15) and the parameters described above. Regions of homology were retained (i.e.,

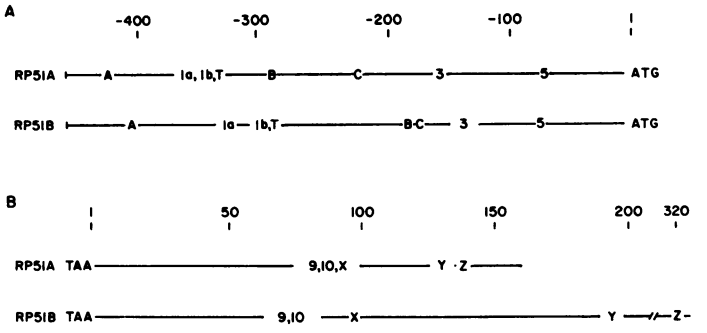

FIGURE 1:
A. Homologies at the 5' ends of the RP51A and RP51B genes. In cases of imperfect matches, the upper sequence corresponds to RP51A and the lower to RP51B.

HOMOL A: TCGAACT

HOMOL B: CCGTTT

HOMOL C: GAGAcGAGG
GAGAtGAGG

1a 1b T
n n_5

HOMOL 1: TAACATCCgTgCATTACAtCCgTACATTTATTTTTTCCA
(a,b,T) TAACATCCaTaCATTACAcCCaTACATTTATTTTTTCCA

n_{11} n_5

HOMOL 3: TgCTTCCT
TcCTTCCT

HOMOL 5: TATTAA

B. Homologies at the 3' ends of RP51A and RP51B.

9 10

HOMOL 9,10: TCtTTAATGTATAaTTAAATAA
TCcTTAATGTATAgTTAAATAA

HOMOL X: AAAATAT

HOMOL Y: GAAGCGTTT

HOMOL Z: TGTAGCT

appear in Figure 1) if they maintained their order within the sequence. Also, a sequence was used only once as a region of homology. Otherwise stated, all homologies were constrained to lie between the same two neighboring homologies (using the

```
A:        -460       -450       -440       -430       -420
        CCCCCATTAT TAATGGAACC TCTGTATTAT ACTTTTCTAT TTCGAACTTT

        -410       -400       -390       -380       -370
        TTGAGACTCA TTCTTGGTAT CCCAGGTGGA CCCAGTAACC TTTTTTCCGG

        -360       -350       -340       -330       -320
        TTTAACATCC GTGCATTACA TCCGTACATT CTATTTTTTA TTTTCCAAAA

        -310       -300       -290       -280       -270
        AACTGGGAGT TCTACTTAAT TTTTTGGCCC CGTTTGGGAA TCTGCTTTGC

        -260
        CACAGGAGGC
```

```
B:                                                      CACA

        -440       -430       -420       -410       -400
        GATAGTAGCA ACATTATAAT CATGGTAATG CAACAGCAAG AGGAAAGTGG

        -390       -380       -370       -360       -350
        AGGGATTAAC GCATTCAGAC AGCTTATAGG GGGAAAGAAA GCAGCAAACT

        -340       -330       -320       -310       -300
        TGCTGCCTGT TCGCAGTCAT TGGTTGCAAA AACTAAACTC TACTCACGCA

        -290       -280       -270       -260       -250
        CACTGGAATG AATGGCAATA TTCTTTTTTA GGTTAACCGG CCG
```

FIGURE 2A: The 5' Upstream DNA Sequence of RP51A. The published sequence of RP51A (6) has been extended by 220 nucleotides to position -460 from the initiating methionine.

FIGURE 2B: The 5' Upstream DNA Sequence of L3. The published sequence of L3 (8) has been extended by 197 nucleotides to position -444 from the initiating methionine by L. Schultz.

homologies found with the composite sequence as initial benchmarks). Identical procedures were adopted for the comparison of the 3' ends of the two RP51 genes.

The published sequences of the ribosomal protein genes RP51A (6) and L3 (8) have been extended (by Abovich and Schultz, respectively) and are shown in Figure 2. The DNA sequences of ribosomal protein genes RP59 (Figure 3) and L16 (Figure 4) were determined by Woolford.

RESULTS

Analysis of Ribosomal Protein Gene DNA Sequences 5' to the Initiation Codon

The DNA sequences upstream from the start methionines of the six *S. cerevisiae* ribosomal protein genes were compared to

```
                                                                    GCTT
-450        -440        -430        -420        -410        -400
ACAAGTTCTG GTATATTCTA TATACTCACT TATTACTTTC AAGTACTTCA CACGGGCCTG
-390        -380        -370        -360        -350        -340
ACATCTCTAT TCTTGTTACT CACTGATTAT CGTTCTTGTT CATACTTGTT ATGTATCTTC
-330        -320        -310        -300        -290        -280
TTTCTCCCTA TTTAAAATGT AATAGAGACT GCTTTGGAGT ACTTACGTGC GGTGTACGGA
-270        -260        -250        -240        -230        -220
CTTAATAAAC ATCTGTACAT TTTACTACTC TTTTTTTTTG CGTTCTTTTT TTCACCTTCA
-210        -200        -190        -180        -170        -160
GCACGAGCAT TGCCATTCTC TACTGCATTT TGGCAAATTG TCTGCTTGCG GCAGACCATC
-150        -140        -130        -120        -110        -100
CGCCTCCTGG CTCATTCCAT ATGGTGCAGG GCTTCCTCAG GTAGACAGTT GAAATGAATT
-90         -80         -70         -60         -50         -40
GGTTATTAAT CACTATATAT TACAGAAAGC ACTTTTTAAT GAAGATTCTA TTTTAAACCC
-30         -20         -10         MET Ser                 20          30
AAGGAGAGTA AGAAACCACA AGAACCCGCC ATG TCT A GTA TGTTTAATCA CATAGTGAAC
        40          50          60          70          80          90
ATTTTTAAAG CATCTCTATT TTCCATTGAT TGTTGTTGAT TGTTTCTGAC GACGTGCAAG
       100         110         120         130         140         150
ATACATTGAA AGTCAGAAAC ATAAAGACAA TTCAACGAAT TCATTGCCTC CAAAGTAATT
       160         170         180         190         200         210
CATAGCGATT AGTTGAGCTT ATTGTGTCAA TGGCAGTATA TTTTGTCAAC TTTTTTTTCG
       220         230         240         250         260         270
ATGGAAAAGC AAAGATACTA TGTAAGAATT AAAAAAAAAA ACTTTTGGAT ACTAACAACA
       280         290         300         310     Asn Val Val Gln Ala
TTACGTTTGA TATCGTCCGA TATCGATTTA CTATTTCCAT TTAG AC GTT GTT CAA GCT

Arg Asp Asn Ser Gln Val Phe Gly Val Ala Arg Ile Tyr Ale Ser Phe Asn
CGT GAC AAT TCC CAA GTT TTT GGT GTT GCT AGA ATT TAC GCT TCT TTC AAC

Asp Thr Phe Val His Val Thr Asp Leu Ser Gly Lys Glu Thr Ile Ala Arg
GAT ACT TTC GTT CAT GTT ACC GAT TTA TCT GGT AAG GAA ACC ATC GCC AGA

Val Thr Gly Gly Met Lys Val Lys Ala Asp Arg Asp Glu Ser Ser Pro Tyr
GTT ACT GGT GGT ATG AAG GTT AAG GCT GAC AGA GAT GAA TCT TCT CCA TAC

Ala Ala Met Leu Ala Ala Gln Asp Val Ala Ala Lys Cys Arg Glu Val Gly
GCT GCT ATG TTG GCT GCC CAA GAT GTT GCC GCT AAG TGT AGG GAA GTC GGT

Ile Thr Ala Val His Val Lys Ile Arg Ala Thr Gly Gly Thr Arg Thr Lys
ATC ACT GCC GTT CAC GTT AAG ATC AGA GCT ACC GGT GGT ACT AGA ACC AAG

Thr Pro Gly Pro Gly Gly Gln Ala Ala Leu Arg Ala Leu Ala Arg Ser Gly
ACT CCA GGT CCA GGT GGT CAA GCT GCT TTG AGA GCT TTG GCC AGA TCT GGT

Leu Arg Ile Gly Arg Ile Glu Asp Val Thr Pro Val Pro Cys Asp Ser Thr
TTG AGA ATT GGC CGT ATC GAA GAT GTT ACC CCA GTT CCA TGT GAC TCC ACC

Arg Lys Lys Gly Gly Arg Arg Gly Arg Arg Leu           730         740
AGA AAG AAG GGT GGT AGA AGA GGT AGA AGA TTA TGA GTTATGCAT GTATTGTACT

       750         760         770         780         790         800
TGTATTGCCG TATTATTTTT TACAGTTAAA AAATGTGTAC ATATAATTAT ATAGCGCCCA
       810         820         830         840         850         860
TAATCAAATC AGCTCATACG TCAATTTAGT AATAAAAAAA AGCCCTTATA ACCTTTTAGT

TAAGAAGA
```

FIGURE 3: DNA Sequence of the S. cerevisiae RP59 Gene and the Inferred Amino Acid Sequence of Ribosomal Protein 59. The initiating and terminator codons, as well as the splicing 5' donor and 3' acceptor sites in the intron have been underlined. DNA sequence established by Woolford.

find regions of homology of a size of 11 bases in which at least 8 (of the 11) bases matched. The length of 5' flanking DNA searched for each gene is shown in Table 1. No matches of 11 consecutive bases were found that are common to all six ribosomal protein gene sequences, yet a 12 base sequence AACATC(T/C)(G/A)T(A/G)CA (HOMOL1, Table 2) was identified that is conserved in at least 5 of the 6 S. cerevisiae ribosomal protein genes. As shown in Table 2, HOMOL1 occurs at a position of about -300 (relative to the start methionine) in the upstream 5'

```
                                                                  TAAAAT
-340        -330        -320        -310        -300        -290
GCAGCAACAT ACATATGTTG AGTTGTATAG ACATCTATAT ATAACAAGCA CAGAACCGTC
-280        -270        -260        -250        -240        -230
TAATTGGTAT TTTTCAGGAC ATTTTAAACA TCCGTACAAC GAGAACCCAT ACATTACTTT
-220        -210        -200        -190        -180        -170
TTTTAATATT CTTTTTGTTT TCATCGCCTT CTTTTTATTT TTATCCGAAG ATCTTTTGGA
-160        -150        -140        -130        -120        -110
ACCCGCTCTG CGAATAGCGA AGCAGGATAC CAAAGTGAAA CTTGGACATA ACTCATCATT
-100        -90         -80         -70         -60         -50
AAAGAAGTAT ACTGTTAAGA GAGGCATTCA TTTCGTGTAT TATAACGTTT AGCATCAGTT
-40         -30         -20         -10         MET Ser Thr Lys Ala Gln
ACCCTTGAAA GCCCAACATA TACAAAAATA CGCGTTCAAG ATG TCT ACT AAA GCC CAA

Asn Pro Met Arg Asp Leu Lys Ile Glu Lys Leu Val Leu Asn Ile Ser Val
AAC CCT ATG CGT GAT TTG AAG ATC GAG AAA TTG GTC TTG AAC ATC TCC GTT

Gly Glu Ser Gly Asp Arg Leu Thr Arg Ala Ser Lys Val Leu Glu Gln Leu
GGT GAA TCT GGT GAC AGA TTA ACC AGA GCC TCC AAG GTT TTA GAA CAA TTA

Ser Gly Gln Thr Pro Val Gln Ser Lys Ala Arg Tyr Thr Val Arg Thr Phe
TCT GGT CAA ACT CCA GTT CAA TCC AAG GCC AGA TAT ACT GTC AGA ACT TTC

Gly Ile Arg Arg Asn Glu Lys Ile Ala Val His Val Thr Val Arg Gly Pro
GGT ATC AGA AGA AAC GAA AAA ATT GCT GTT CAC GTT ACC GTC AGA GGT CCA

Lys Ala Glu Glu Ile Leu Glu Arg Gly Leu Lys Val Lys Glu Tyr Gln Leu
AAG GCT GAA GAA ATT TTG GAA AGA GGT TTG AAG GTC AAG GAA TAC CAA TTG

Arg Asp Arg Asn Phe Ser Ala Thr Gly Asn Phe Gly Phe Gly Ile Asp Glu
AGA GAC AGA AAC TTC TCT GCT ACC GGT AAC TTC GGT TTC GGT ATT GAC GAA

His Ile Asp Leu Gly Ile Lys Tyr Asp Pro Ser Ile Gly Ile Phe Gly Met
CAC ATT GAC TTG GGT ATC AAG TAT GAC CCA TCC ATC GGT ATT TTC GGT ATG

Asp Phe Tyr Val Val Met Asn Arg Pro Gly Ala Arg Ala Thr Arg Arg Lys
GAT TTC TAT GTC GTC ATG AAC AGA CCA GGT GCT AGA GTC ACT AGA AGA AAG

Arg Cys Lys Gly Thr Val Gly Asn Ser His Lys Thr Thr Lys Glu Asp Thr
AGA TGT AAG GGT ACT GTT GGT AAC TCC CAC AAG ACA ACT AAG GAA GAC ACC

Val Ser Trp Phe Lys Gln Lys Tyr Asp Ala Asp Val Leu Asp Lys       530
GTC TCT TGG TTC AAG CAA AAG TAC GAC GCT GAT GTG CTC GAT AAA TAA TTTGG
      540         550         560         570         580         590
TCTCGGTATA GTCAGTGACA ACATCAACTA CTTAATATAT AAGAACAAAT AAAATATCCC
      600         610         620         630         640         650
AAAAATATCA TATATCCTCA TCACATTTGC AAGTCTAGCG CTTCGATGCG TTGTGAACAC
      660         670         680         690         700         710
TTTGTCAATG TATTTAGTTG TATTCATACC CAATTTATTG GCACTTATTT GATACTCACC
      720         730         740         750         760         770
ATGCAGGGTA ATAGAAAATG TGCTGAAAAA AAGCTAAACC TTTCTTATTA AGAAAATGGG
      780         790         800         810         820         830
AACCATAACA GTGGTTCGAT TAATGAGGGA CCAATACTGT TGATAAGGGC ATTGCACCGA
      840         850         860         870         880         890
GCAACGACCA ACAAGAAAAT GTTCAGAAGT ACAGTTTGGA GACGTTTTGC ATCTACCGGC
      900         910         920         930         940         950
GAAATTGCGA AAGCAAAGCT GGATGAATTC TTGATATACC ACAAGACAGA TGCGAAACTA
      960         970         980         990        1000        1010
AAACCATTCA TTTACCGTCC CAAGAATGCT CAGATATTGT TAACTAAAGA TATTAGGGAT
     1020        1030        1040        1050
CCAAAAACAA GGAACCATTA CAACCGAGAC CTCCCGTAAG
```

FIGURE 4: DNA Sequence of the S. cerevisiae L16 Gene and the Inferred Amino Acid Sequence of Ribosomal Protein L16. The initiating and terminator codons have been underlined. DNA sequence established by Woolford.

flanking DNA of ribosomal protein genes RP51A, RP51B, L29, RP59, and L16, as well as in the upstream region of the S. carlsbergensis gene encoding L17a. The conserved sequence is identical in L29, RP59, and L17a, whereas in L16, RP51A and RP51B the sequence varies at one or more positions. The ribosomal protein genes RP51A and RP51B each contain a sequence

TABLE I: DNA Sequences.

	Sequence analysed: 5' to initiation codon	3' to termination codon	Source
Ribosomal Protein Genes			
S. cerevisiae			
RP51A	460 bp	239 bp	(6) and Figure 2A
RP51B	510 bp	390 bp	(34)
RP59	454 bp	149 bp	Woolford, Figure 3
L29	322 bp	97 bp	(7)
L16	346 bp	524 bp	Woolford, Figure 4
L3	444 bp	120 bp	(8) and Figure 2B
S. carlsbergensis			
L17a	433 bp	206 bp	(36)
S10	146 bp	89 bp	(9)
Non-Ribosomal Protein Genes			
S. cerevisiae			
ADH1	1390 bp	667 bp	(10)
ENOA	353 bp	347 bp	(11)
ENOB	180 bp	364 bp	(11)
HIS1	1190 bp	729 bp	(12)
HIS4	1332 bp	1020 bp	(13)
MATA1	1534 bp	450 bp	(14)
MATA2	1193 bp	878 bp	(14)

matching HOMOL1 (HOMOL1a, Table 2) and also contain a second sequence similar to HOMOL1 (HOMOL1b, Table 2) which varies slightly from the consensus sequence. In both RP51A and RP51B, HOMOL1b occurs within 15 bp downstream from HOMOL1a (Figure 1). The consensus sequence HOMOL1 was not found in the available upstream region (see Table 2) of the *S. cerevisiae* gene L3 or in the *S. carlsbergensis* gene S10. It should be noted that we have

TABLE II: Homologous sequences 5' to the Initiation Codon.

	HOMOL1	HOMOL2	HOMOL3
CONSENSUS	AACATC$^{TG}_{CA}$T$^{A}_{G}$CA	$^{T}_{A}$CATCTNTA	$^{TC}_{GG}$CTTC$^{T}_{C}$T
RP51A	AACATCCGTGCA (-357)a	N.F.	TGCTTCCT (-150)
	tACATCCGTACA (-344)b		
RP51B	AACATCCATACA (-323)a	N.F.	TCCTTCCT (-128)
	tACAcCCATACA (-299)b		
L29	AACATCTGTACA (-279)	ACATCTGTA (-278)	TCCTTCTT (-57)
		TCATCTGTA (-108)	
RP59	AACATCTGTACA (-263)	ACATCTGTA (-262)	GGCTTCCT (-121)
		ACATCTCTA (-390)	
L16	AACATCCGTACA (-254)	ACATCTATA (-310)	GCCTTCTT (-195)
L3	N.F.	TCATCTCTA (-38)	TGCTTCCT (-124)
S10	N.F.	TCATCTTTA (-102)	N.F.
L17a	AACATCTGTACA (-351)	ACATCTGTA (-350)	N.F.

	HOMOL4	HOMOL5	HOMOL6
CONSENSUS	$^{T}_{A}$AT$^{T}_{A}$TTNCA	TATT$^{TT}_{AA}$	TCAAGA
RP51A	TATTTTCCA (-322)	TATTAA (-66)	TCAAGA (-28)
RP51B	TATTTTACA (-78)	TATTAA (-69)	TAAAGA (-21)
L29	TATTTTCCA (-134)	TATTTA (-92)	TCAAGA (-6)
RP59	TATATTACA (-75)	TATTTT (-42)	ACAAGA (-13)
L16	TATTTTTCA (-273)	TATTAT (-63)	TCAAGA (-5)*
L3	AATTTTTCA (-146)	TATTTA (-97)	TCTAGA (-33)
S10	N.F.	N.F.	N.F.
L17a	TATTTTTCA (-303)	TATTTT (-60)	ACAAGX (-40)

* Last A = first A of ATG

N.F.= Not Found

limited 5' flanking DNA sequence information for the S10 gene (see discussion), as well as some non-ribosomal protein yeast genes, so that HOMOL1 may be upstream of what has been sequenced. Another sequence (T/A)CATCTNTA (Table 2, HOMOL2) which is very similar to HOMOL1 was also identified. It is not found in RP51A and RP51B and does not appear to be conserved with

respect to its position within the 5' region of those ribosomal protein genes that contain it. In ribosomal protein genes L29, RP59, and L17a, HOMOL1 is included as a subset of the matches to HOMOL2.

A third consensus sequence (T/G)(C/G)CTTC(T/C)T (Table 2, HOMOL3), was identified in the 5' ends of the S. cerevisiae ribosomal protein genes. This sequence is found at a distance of -120 to -200 (with respect to the start methionine) in each ribosomal protein gene, with the exception of L29 (in which the sequence occurs at -57). This sequence was not found in the S. carlsbergensis ribosomal protein genes, or in most of the non-ribosomal protein genes that were examined. (HOMOL3 does occur in ADH1 at position -139 and also occurs in the coding region of the genes ENO8 and HIS4.)

Another sequence found only in the 5' region of S. cerevisiae ribosomal protein genes is the AT rich sequence (T/A)AT(T/A)TTNCA (Table 2, HOMOL4), which does not appear to be conserved with respect to its position relative to the start methionine. Also, an AT rich sequence resembling a "TATA" box (16) was identified for each of the ribosomal protein genes (Table 2, HOMOL5). The sequence TATT(T/A)(T/A) was found within a range of -42 to -97 for 7 out of the 8 ribosomal protein genes (although most of the ribosomal protein genes contain additional matches to this sequence in their more distal 5' regions). This consensus sequence was not found in the region preceding the gene for S10 (although the AT sequence TTAATT occurs in S10 at -40 and -80).

In an attempt to identify homologies in regions that are likely to be transcribed and represent untranslated leader sequences, the sequence directly preceding the start methionine in each gene was compared by eye. A short homology TCAAGA (Table 2, HOMOL6) was identified at a postion of -6 with respect to the start methionine in the ribosomal protein genes L16 and L29 (in both cases the sequences occur between the start methionine and the transcription start). This sequence is also found in the 5' untranslated region of the RP51A gene at -28 (and is thus included in the longer of the two major 5' termini that occur in this region of RP51A mRNA (6)). A sequence differing from TCAAGA

at one nucleotide occurs within the first 50 nucleotides preceding the start methionine for the ribosomal protein genes L3, RP51B, and RP59. (In each case, the single base change involves only the first three nucleotides: T, C, or A.) The sequence TCAAG also occurs in the 5' untranslated region of non-ribosomal protein yeast genes such as ADH1 and PKG, and it has been suggested that this sequence is involved in the initiation of transcription of highly expressed genes in yeast (17, 18).

The 5' regions were also examined for open reading frames. No open reading frames larger than 34 amino acids could be identified in the 5' flanking DNA (within the extent of the available sequence analyzed for each gene, Table 2) of any of the S. cerevisiae ribosomal protein genes except L16, in which an open reading frame of 97 amino acids occurs (which terminates at TAG, -48 bp preceding the L16 start methionine). This open reading frame is largely composed of codons which are rare in yeast (data not shown; 10).

Analysis of Ribosomal Protein Gene Introns

It is striking that most of the cloned ribosomal protein genes which have been analyzed contain a single intron located at the 5' end of the gene (6, 7, 9, 19, 20, 36; Woolford, J., Larkin, J. and Levy, A., Figure 3). The yeast actin gene and the matal gene are currently the only non-ribosomal protein genes in yeast reported to have introns (21, 22; 35). The intron splice junctions of yeast conform to the consensus splice sequence determined from examination of metazoan mRNA splice junctions (23, 24). The 5' splice sites are strikingly conserved for each of the spliced yeast genes, while the 3' splice sites show more variability (Table 3).

The S. cerevisiae ribosomal protein gene introns were compared in a manner similar to the 5' end comparison described above. The largest complete homology between the four introns corresponds to the seven base ICS (intron conserved sequence) TACTAAC (Table 3, HOMOL7) which has been previously described (25, 26). This sequence is identical in all yeast introns (including the non-ribosomal protein genes actin and matal), and is found within -65 bp of the 3' splice site in every case. No other 7/7 base homologies were found in all six ribosomal pro-

tein gene introns.

When the four S. cerevisiae ribosomal protein gene introns (RP51A, RP51B, L29 and RP59) were compared as a group, an additional sequence was found T(T/C)NCATTT(G/A) (Table 3, HOMOL8), which is not found in S. carlsbergensis ribosomal protein gene introns (L17a and S10) or in non-ribosomal protein gene introns (actin and matal). This sequence is not conserved with respect to its position within the intron, and is found once in each intron with the exception of L29 (where it is found twice). HOMOL8 does not occur within the coding region of any ribosomal protein gene or within the untranslated regions of any ribosomal protein mRNA. In the introns of L29 and RP59, HOMOL8 is part of a larger homology in which 10/11 bases are matched. When the S. carlsbergensis ribosomal protein gene introns (L17a and S10) were compared as a group, a different sequence A(T/A)TCTAATG(G/A)T was identified as being common to both S. carlsbergensis ribosomal protein gene introns, yet absent from the S. cerevisiae introns (including ribosomal protein gene introns and non-ribosomal protein gene introns). Open reading frames were searched for within each intron; none were longer than 117 nucleotides.

Analysis of Ribosomal Protein Gene DNA Sequence 3' to the Termination Codon

The analysis of ribosomal protein DNA sequences which occur 3' to the termination codon was carried out in the same way as has been described for the 5' regions and introns. No sequence matches greater than five consecutive base pairs could be identified as being found at the 3' ends of all eight ribosomal protein genes. A sequence was identified, however, TNNATGTAT (Table 3B, HOMOL9), which is found in all S. cerevisiae ribosomal protein genes within a distance of 130 bases from the termination codon. This sequence was also identified in some non-ribosomal protein S. cerevisiae genes examined.

The sequence AAUAAA is known to be involved in the addition of poly A tails to mRNA in higher eucaryotes (27). DNA sequences 3' to the termination codon were therefore searched for the sequence AATAA (Table 3B, HOMOL10). The sequence AATAA (HOMOL10) was found within 110 bp of the stop codon of each

TABLE III. A. Homologous Sequences Within Introns.

	HOMOL7	HOMOL8	5' splice junction	3' splice junction
CONSENSUS	TACTAAC	$T^{T}_{C}NCATTT^{G}_{A}$		
RP51A	TACTAAC (-64)	TTGCATTTA (-49)	GTATGT	TAATAG
RP51B	TACTAAC (-43)	TCTCATTTG (-28)	GTACGT	TTATAG
L29	TACTAAC (-49)	TTCCATTTG (-507)	GTATGT	GTACAG
		TTTCATTTA (-176)		
RP59	TACTAAC (-53)	TTCCATTTA (-10)	GTATGT	ATTTAG
S10	TACTAAC (-28)	N.F.	GTATGT	TAACAG
L17a	TACTAAC (-50)	N.F.	GTATGT	AACTAG
Actin*	TACTAAC (-49)	N.F.	GTATGT	GTTTAG
MATa1* (1)	TACTAAC (-17)	N.F.	GTATGT	CTTCAG
(2)	TACTAAC (-16)	N.F.	GTATGT	TTGTAG

B. Homologous Sequences 3' to the Termination Codon.

	HOMOL9	HOMOL10
CONSENSUS	TNNATGTAT	AATAA
RP51A	TTAATGTAT (+81)	AATAA(T) (+24)
		AATAA(A) (+92)
RP51B	TTAATGTAT (+66)	AATAA(T) (+80)
L29	TCTATGTAT (+18)	AATAA(A) (+32)
RP59	TGCATGTAT (+5)	AATAA(A) (+110)
L16	TCAATGTAT (+130)	AATAA(A) (+53)
L3	TTTATGTAT (+77)	AATAA(A) (+16)
S10	N.F.	N.F.
L17a	N.F.	AATAA(C) (+60)

Positions of HOMOL7 and HOMOL8 are shown as the nucleotide distance relative to the 3' splice junction. HOMOL9 and HOMOL10 nucleotide positions are relative to the termination codon. N.F.= Not Found

* Non-ribosomal protein yeast genes which contain introns.

ribosomal protein gene, except for the gene S10.

Analysis of Coding Regions

The coding regions of the eight ribosomal protein genes were assembled in a composite file (without intervening

sequences) and analyzed in terms of both protein sequence and DNA sequence. Codon frequency was calculated for the composite file, and also for individual ribosomal protein gene coding regions. Codon usage in ribosomal protein genes (when compared individually and as a group) is similar, both quantitatively and qualitatively, to other highly expressed genes in yeast (data not shown) (28). Non-preferred codons are used only occasionally by all eight ribosomal protein genes.

Analysis of RP51A and RP51B

The DNA sequences encoding RP51A and RP51B were compared to identify DNA sequences which are conserved between the two RP51 genes. Both genes are transcribed in vivo and the two genes complement. The two proteins differ at only one of 133 amino acids and the coding DNA is 96% homologous (34). In contrast, the DNA sequences flanking the coding DNA are largely non-homologous (see below). Since both genes are functional ribosomal protein genes and probably diverged from a common ancestor, conserved sequences, in addition to those identified above, might function in the expression of ribosomal protein synthesis in general or of RP51 synthesis in particular.

In a comparison of the sequences 5' to the initiator methionine, four additional homologies were identified (Figure 1A). HOMOL A occurs approximately -420 5' to the initiator methionine in both RP51A and RP51B. The RP51A DNA sequence containing HOMOL A (and also HOMOL1a,b) extends beyond the previously reported RP51A sequence (6) and is shown in Figure 2. HOMOL A is quite well conserved with respect to position and is the most upstream homology element identified in this study. The region denoted by T (for a T-rich stretch) has been combined with regions 1a and 1b to generate one large conserved region, 1a-1b-T. HOMOL4 is not identified in Figure 1 because it lies upstream of HOMOL3 in RP51A and downstream of HOMOL3 in RP51B (see Methods and Table II).

In a comparison of the DNA sequences occurring 3' to the termination codon of each RP51 gene (Figure 1B), HOMOL9 and HOMOL10 were found to be included within a larger homologous region of 22 bp (occurring about 75 bp downstream from the termination codon). Three additional homologies were identified, of

which HOMOL Y (GAAGCGTTT) (at a position 126 bp and 189 bp downstream from the termination codon in RP51A and RP51B, respectively) is the most extensive.

DISCUSSION

Although the DNA sequence is known for many yeast genes, little is known of specific DNA sequences having regulatory properties. Deletion analysis of 5' DNA sequences for the genes His3 (29), CYC1 (30), and ADH1 (18) have in each case identified DNA regions likely to be involved in regulation specific to a particular gene. In each case DNA sequences relatively far upstream from the start of transcription are required for function (the actual distance ranges between 120-350 bp depending on the gene; 31). Genes which are regulated in common presumably share DNA sequence homology by which regulation can be selectively directed. In the case of genes which are regulated in common by general amino acid control (His1, His3, His4, Trp5), the sequence A(A/T)GTGACTC has been identified, occurring in repeats at a variable distance from the start methionine within the 5' ends of these genes (12). Similarly, the sequence AA(C/A)A(C/A)CCAAG occurs at a position approximately -30 to -60 in the region of five genes encoding glycolysis functions (32). Not surprisingly, these sequences do not occur in the 5' regions of ribosomal protein genes, yet they do provide an indication of the extent to which other commonly regulated genes share sequence homology. It is interesting to note that of the consensus sequences that have been identified, the number of invariant bases (underlined above) that comprise the consensus sequence is small. In the case of the general amino acid control consensus sequence A(A/T)GTGACTC, only three of the four invariant bases are colinear, suggesting that some sequences involved in gene regulation may be difficult to identify by a computer analysis of DNA sequences (as the signal/noise ratio would be quite poor at this criterion). Nevertheless, a comparison of yeast ribosomal protein gene DNA sequences was undertaken, given that no alternative data exist at present (such as deletion mapping analysis) from which to predict DNA sequences likely to be involved in the coordinate expression of ribosomal protein genes.

The most notable 5' region homology identified corresponds to HOMOL1, AACATC(T/C)(G/A)T(A/G)CA which is found in RP51A, RP51B, L16, RP59, and L29, and is conserved with respect to the sequence and its approximate position relative to the start methionine. In all cases (see Figure 1 for an example) HOMOL1 is followed by a stretch of T residues. HOMOL1 is an attractive candidate for a promoter or enhancer element specific to the transcriptional control of ribosomal protein genes, as it was not found in the 5' regions of any non-ribosomal protein genes examined. It is interesting to note that the sequence is also present, at the same location, in the S. carlsbergensis gene L17A. Although HOMOL1 was not found in S10, a preliminary analysis of additional upstream S10 sequence data, suggest that this ribosomal protein gene also contains this sequence element (R. Leer, personal communication). This is consistent with the notion that S. carlsbergensis is similar to S. cerevisiae (R. Planta, personal communication). An examination of the DNA sequence of L3 to -444 suggests that HOMOL1 may be absent from this gene. If so, it may be associated with many but not all ribosomal protein genes.

A preliminary statistical analysis suggests that the presence of HOMOL1 upstream of most of the ribosomal protein genes is highly significant (data not shown). In contrast, many of the other identified homologies are considerably less significant, suggesting that some of these (e.g., HOMOL5) may be fortuitous rather than due to functional constraints. Experimental data will be required to determine whether HOMOL1 (and also HOMOL2-6) play a role in ribosomal protein gene expression.

For those ribosomal protein genes that contain introns (Table 3), the 5' splice site sequence is clearly conserved. Although the GT of the 5' splice site is very highly conserved in metazoan 5' splice sites (PuGTXXG), there is generally some variation in the remaining nucleotides of the consensus sequence. In yeast there is almost no variation from the 5' sequence GTATGT (RP51B is one exception, having GTACGT; S10B is another, having GTATGA, Planta, R., personal communication). The

sequence specificity of the 5' splice site in yeast thus appears to be more stringent than in other eukaryotes. The 3' splice sites of yeast show variation characteristic of metazoan 3' sites (PyPyNPyAG)(24).

Within the introns of ribosomal protein genes, the most notable homology corresponds to the sequence TACTAAC (HOMOL7), which occurs in all sequenced yeast nuclear mRNA introns. This sequence is likely to be involved in the splicing of all introns in yeast (26, 25) and not related to specific regulation of ribosomal protein mRNA levels. A sequence specific to ribosomal protein introns was identified [T(T/C)NCATTT(G/A) HOMOL8], yet no experimental data exist at this time to suggest that this sequence is involved in regulation. Similarly the significance of HOMOL9 (TNNATGAAT) and HOMOL10 (AATAA), occurring in the 3' DNA sequence of ribosomal protein genes, remains speculative.

With the non-stringent parameters used in the comparisons presented in this report, a two gene analysis always leads to a larger number of homologies than a multigene analysis requiring a match in more than two genes (Figure 1, Table II and our unpublished data). In this context, it is of interest that a significant fraction of the homologies shown in Figure 1 are also present in Table II, suggesting that these regions (HOMOL1,3,5,9,10) are of general significance to ribosomal protein expression or regulation. Neither HOMOL2 nor HOMOL4 is well conserved with respect to position, leading to the suggestion that they are not significant. Perhaps HOMOL2 is due to its similarity to HOMOL1, and HOMOL4 is due to AT-rich DNA. (HOMOL4 is not present in Figure 1 because it lies upstream of HOMOL3 in RP51A and downstream of HOMOL3 in RP51B.) HOMOL5 is probably the ribosomal protein gene "TATA box." HOMOL6 and HOMOL3 are short and difficult to evaluate (although HOMOL3 is quite well conserved with respect to position, with the exception of L29). The homologies which are RP51-specific are also difficult to evaluate. HOMOL A is far upstream of the coding regions. If it exists in other ribosomal protein genes, we may not have sufficient upstream sequence to identify it elsewhere. Alternatively, HOMOL A and other RP51-specific homologies may be due to neighboring genes, upstream in the case

of A,B,C and downstream in the case of X,Y,Z.

As described above, our data do not exclude the presence of other homologies not shown in Figure 1 or Table 2. It would appear, however, that HOMOL1 and 3, especially HOMOL1 and its adjacent downstream T-rich segment, are the best candidates for ribosomal protein gene specific elements. Indeed, analysis of preliminary DNA sequences of a ribosomal protein gene not included in this study suggests that HOMOL1 is a feature associated with at least one additional S. cerevisiae yeast ribosomal protein gene (33). The significance of HOMOL1, as well as other homologies specific to ribosomal protein genes, will be clarified as new ribosomal protein gene DNA sequences and experimental data become available.

ACKNOWLEDGEMENTS

The authors for the Rosbash laboratory (J.L.T., N.A., and M.R.) would like to thank Jim Pustell for his advice and help on the computer analysis and Tobie Tishman for preparing the manuscript.

[2]Albert Einstein College of Medicine of Yeshiva University, 1300 Morris Park Avenue, Bronx, NY 10461, USA

[3]Department of Biological Sciences, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, PA 15213, USA

[4]Biochemish Laboratorium, Vrije Universiteit, 10091 HV Amsterdam, The Netherlands

[5]Department of Medical Genetics, University of Toronto, Toronto, Ontario, Canada

[6]School of Medicine, Department of Biochemistry and Nutrition, University of North Carolina at Chapel Hill, Faculty Laboratory Office Building 231H, Chapel Hill, NC 27514-7231, USA

REFERENCES

1. Gorenstein, C. and Warner, J.R. (1976) Proc. Natl. Acad. Sci. USA 73, 1547-1551.
2. Warner, J.R. and Gorenstein, C. (1977) Cell 11, 201-212.
3. Faliks, D. and Meyuhas, O. (1982) Nucl. Acids Res. 10, 789-801.
4. Kim, C.H. and Warner, J.R. (1983) Molec. and Cell. Biol. 3, 457-465.
5. Pearson, N.J., Fried, H.M. and Warner, J.R. (1982) Cell 29, 347-355.
6. Teem, J. and Rosbash, M. (1983) Proc. Natl. Acad. Sci. USA. 80, 4403-4407.

7. Kaufer, N.F., Fried, H.M., Schwindinger, W.F., Jasin, M. and Warner, J.R. (1983) Nucl. Acids Res. 11, 3123-3135.
8. Schultz, L., and Friesen, J. D. (1983) J. Bact. 155, 8-14.
9. Leer, R. J., van Raamsdonk-Duin, M.M.C., Molenaar, C.M.Th., Cohen, L.H., Mager, W.H. and Planta, R.J., Nucl. Acids Res. 10, 5869-5878.
10. Bennetzen, J.L. and Hall, B.D. (1982) J. of Biol. Chem. 257, 3018-3025.
11. Holland, M.J., Holland, J.P., Thill, G.P. and Jackson, K.A. (1981) J. of Biol. Chem. 256, 1385-1395.
12. Hinnebusch, A.G. and Fink, G.R. (1983) J. of Biol. Chem. 258, 5238-5247.
13. Donahue, T.F., Farabaugh, P.J., Fink, G.R. (1982) Gene 18, 47-59.
14. Astell, C.R., Ahlstrom-Jonasson, L. and Smith, M. (1981) Cell 27, 15-23.
15. Pustell, J. and Kafatos, F.C. (1982) Nucl. Acids Res. 10, 4765-4781.
16. Benoist, C. and Chambon, P. (1981) Nature 290, 304-310.
17. Dobson, M. J., Tuite, N. A., Kingsman, A. J., and Kingsman, S.M. (1982) Nucl. Acids Research 8, 2625.
18. Beier, D.R. and Young, E.T. (1982) Nature 300, 724-728.
19. Bollen, G.H.P.M., Molenaar, M.R., Cohen, L.H., van Raamsdonk-Duin, M.M.C., Mager, W.H. and Planta, R.J. (1982) Gene 18, 29-38.
20. Fried, H.M., Pearson, N.J., Kim, C.H., and Warner, J.R. (1981) J. Biol. Chem. 259, 10176-10183.
21. Gallwitz, D. and Sures, I. (1980) Proc. Nat. Acad. Sci. USA 77, 2546-2550.
22. Ng, R. and Abelson, J. (1980) Proc. Nat. Acad. Sci. USA 77, 3912-3916.
23. Breathnach, R. and Chambon, P. (1981) Ann. Rev. Biochem. 50, 349-383.
24. Mount, S.M. (1982) Nucl. Acids Res. 10, 459-472.
25. Langford, C., J. and Gallwitz, D. (1983) Cell 33, 519-527.
26. Pikielny, C., Teem, J.L. and Rosbash, M. (1983) Cell 34, 395-403.
27. Fitzgerald, M. and Shenk, T. (1981) Cell 24, 251-260.
28. Bennetzen J. L. and Hall, B. D. (1982) J. Biol. Chem. 257, 3026-3031.
29. Struhl, K. (1981) Nature 305, 347-458.
30. Guarente, L. and Ptashne, M. (1981) Proc. Natl. Acad. Sci. USA 78, 2199-2203.
31. Struhl, K. (1983) Nature 305, 347-458.
32. Holland, J.P., Labieniec, L., Swimmer, C. and Holland, M.J. (1983) J. of Biol. Chem. 258, 5291-5299.
33. Mitra, G. and Warner, J.R. (1984) J. Biol. Chem. In press.
34. Abovich, N. and Rosbash, M.R. (1984) Molec. and Cell. Biol. 4, 1871-1879.
35. Miller, A.M. (1984) The EMBO J. 3, 1061-1066.
36. Leer, R.J., Van Raamsdonk-Duin, M.M.C., Hagendoorn, M.J.M., Mager, W.H. and Planta, R.J. (1984) Nucl. Acids Res. 12, 6685-6700.