Nucleotide sequence of a yeast Ty element: Evidence for an unusual mechanism of gene expression

(DNA sequencing/protein homology/S1 nuclease mapping/ β -galactosidase fusion/translational frameshifting)

JEFFREY CLARE AND PHILIP FARABAUGH

Department of Microbiology, University of Connecticut Health Center, Farmington, CT 06032

Communicated by Gerald R. Fink, December 10, 1984

ABSTRACT We have determined the DNA sequence of the transposable element Ty912 of yeast. The 5918-base-pair element encodes two genes, tya912 and tyb912, which specify proteins similar to sequence-specific DNA-binding proteins of Escherichia coli and retroviral reverse transcriptases, respectively. The tyb912 gene is atypical of eukaryotic genes since (i) it begins 1336 nucleotides into the Ty912 mRNA (i.e., downstream of the tva912 gene) and (ii) the first in-frame AUG is 921 nucleotides into the coding frame. Protein blot analysis of Ty-lacZ fusions shows that the tyb912 gene is translated starting at the 5' end of the tya912 gene and that the primary translational product is a tya912::tyb912 fusion protein. We have shown that synthesis of this fusion protein probably does not occur by RNA splicing. The data are consistent with a mechanism of translational frameshifting occurring within the region of overlap between the 3' end of tya912 and the 5' end of tyb912.

Ty elements are a family of approximately 30 transposable elements, which are dispersed throughout the yeast genome (1-4). Details of the structure of Ty elements suggest that they are more closely related to the retroviruses of higher cells than to bacterial transposons. They consist of a central region (termed epsilon) approximately 5300 base pairs (bp) in length surrounded by 330-bp delta elements. Elder et al (5) showed that the major Ty-encoded mRNA has a terminal repeat similar to the R region of retroviral provirus transcripts. At the 5' delta-epsilon boundary of several Ty elements there is a region complementary to the 3' end of the initiator methionine tRNA of yeast (4), a structure analogous to the negative strand primer binding site of retroviruses. Varmus (6) has pointed out that Ty elements also include an oligopurine tract adjacent to the 3' delta that is similar to the binding site for the retrovirus primer of positive strand synthesis. These factors, all of which are required for the retroviral life cycle, imply that Ty elements are related to metazoan retroviruses. In this report we present further evidence to support this idea, both from the complete DNA sequence of a Tyl element and from studies of the unusual organization and expression of its gene products.

MATERIALS AND METHODS

DNA Manipulations. To construct the tyb912::lacZ fusion plasmid, pPF350, a 2.64-kilobase-pair (kb) Sal I/Bgl II fragment containing the 5' end of Ty912 was ligated into Xho I/BamHI-digested pLGSD5, a high copy number lacZ fusion vector based on the 2- μ m element (7). This procedure gives an in-frame fusion of lacZ to the Bgl II site of tyb912 at nucleotide (nt) 1702. Plasmid pJC8 was made by opening pPF350 at the unique Sph I site at nt 1218 and removing the cohesive ends with T4 DNA polymerase, followed by ligating the blunt ends. Plasmids used for sequencing by the Maxam and Gilbert method (8) were Sal I and Pst I subclones of YIp312, previously described (3).

S1 Nuclease Analysis. To map Ty transcripts with S1 nuclease, total RNA was prepared as described (5) from a pPF350 yeast transformant and enriched for poly(A)⁺ RNA by using oligo(dT)-cellulose chromatography. The DNA RNA hybrids protected from S1 nuclease digestion were separated on neutral and alkaline agarose gels (9) and visualized by Southern hybridization. Under these conditions a splice that removes as little as 11 nt should be detected (9). The *lacZ*-specific Southern probe was a 0.75-kb *Bam*HI/*Cla* I restriction fragment from pLGSD5. The Ty912-specific probe used was a 2.7-kb *Bgl* II fragment isolated from a YIp312 subclone. This fragment contains the complete 5' region of the element up to the *Bgl* II site at nt 1702.

Isolation of tyb912::lacZ Fusion Proteins. Fusion proteins were isolated from a yeast transformant of pPF350 by affinity chromatography using p-aminobenzyl-1-thio- β -galactosamine-agarose (10). Proteins retained by the column were analyzed by electrophoresis in 9% NaDodSO₄ polyacrylamide gels and blotting (11), using monoclonal mouse IgG anti- β -galactosidase (a gift from J. Partaledis). Bands were visualized by using goat anti-mouse IgG antibody conjugated with horseradish peroxidase and Bio-Rad HRP color development reagent.

β-Galactosidase Assays. Protein extracts were assayed for β-galactosidase activity as described by Miller (12). β-Galactosidase activities of growing yeast cultures were determined by the rapid chloroform/NaDodSO₄ permeabilization procedure (7).

RESULTS

The Ty912 Element Includes Two Open Reading Frames. We have sequenced the Ty912 element, which inserted into the promoter of the his4 gene, causing the his4-912 mutation (3, 13). The element is 5918 bp in length and consists of a 5250-bp region flanked by a 334-bp perfect direct repeat called *delta912*. The sequence in Fig. 1 is the coding strand of Ty912. Inspection of the Ty912 sequence reveals two open reading frames covering 89% of the element (5269 bp). The 5'-proximal frame, tya912, which begins at an ATG codon at nt 294 and continues to a TGA codon at nt 1614, could encode a protein of 440 amino acids with a predicted molecular weight of 48,900. The AUG nearest the 5' end of the Ty mRNA is the first AUG in tya912, so we would expect the gene to be translated efficiently in vivo. Furthermore, the context around this AUG, A-C-A-A-U-G-G, favors initiation (16). A second frame, tyb912 overlaps the last 38 bp of tya912 and extends from an ACA codon at nt 1576 to a TAG codon at nt 5560. The first ATG codon in tyb912 is 921 bp into the gene at nt 2497. It is likely, given its length, that the

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. \$1734 solely to indicate this fact.

Abbreviations: bp, base pair(s); kb, kilobase pair(s); nt, nucleotide(s); RSV, Rous sarcoma virus.

2830 Genetics: Clare and Farabaugh

i	TETTEBAATAGAAATCAACTATCATCTACTAACTAETATTACCATTACTAETATTATCATATACEETETTABAAGATBACECAAATGATBASAAATABTCATCTAAATTAETBBAAGECBAAACECAAGBATTBATAATBTAATABBA
151	Netéjuser <u>TCAATGAATATAAACATATAATAATGATGATAATAATATTATAGAATTGTGTAGAATTGCAGATTCCCTTTATGGATTCCTAAATCCTTGGTATAATGTCTGTATAACATGGCTTTATGGCCTTTATGGAATGGGAT TCAATGAATAACATATAACATAGTATGATGATGATGATGATGATGTGTGGAGATTGCAGGATTCCCTTTATGGATTCCTGAAGCATGGGAGAACTTCTGGTATATCTGTAT</u>
301	Start of tysyl2 61a61aLeuSer61aNisSerProIleSerNis61ySerA1aCysA1aSerValThrSerLys61uValNisThrAsa61aAspProLeuAspValSerA1aSerLysThr61u61uCys61uLysValSerThrLysA1aAsaSer61a <u>CCCAACAATTATCTCAACATTCACCCATTCTCA</u> T66TA6C6CCT6T6CTTC66TTACTTCTAA66AA6GTCCCACAAAATCAA6ATCCSTTA8AC6TTCG6CTTCCGACAAAGAA6AATBT6A6AA66TCCTCCACTAA6CCTAACTCC
451	61aThrThrThrThrAroA1aSerSerA1aVa1ProB1uAsaProHisHisA1aSerProB1aProA1aSerVa1ProProLeuB1aAsaB1yProTyrProB1aB1aCysHetHetThrB1aAsaB1aAsaProSerB1yTrpSerPhe AACABACAACAACACCACCTGCTTCATCABCTGTTCCAABAGAACCCCCCATCATGCCTCCTCCAACCTGCTTCAGTACCCACCTTCAGAAAGGGAGTGCATGCA
601	Tyr&lyHisProSerNetIleProTyrThrProTyrGlaNetSerProNetTyrPheProPro&lyPro&laSer&laPhePro&laTyrProSerSerVal&lyThrProLeuSerThrProSerPro&luSer&lyAsaThrPheThr TTTAC68ACACCCATCTAT6ATTCC6TATACACCTTAT6AAAT6T66CCTAT6TACTTTCCACCT68BCCACAATCACA6TTTCC6CA6TATCCATCATCAGTT86AAC6CCTCT8A8CACTCCATCACCT8A8TCAA8TAACATTTA
751	AspSer Ser AlaAspSer AspNetThr Ser Thr LysLysTyr Val Ar gProProProNetLeuThr Ser Pro Asp AspPhePro AspTr pValLysThr Tyr IieLysPheLeu®ib AspSer AspLeu®iy@iyIieIieProThr CTGATTCATCCTCAGCGGGGCTCTGATAGACAATAGCCAACGAAAAATATGTCAGACCACCGACGTTAGCCTCACCGAATGGCTTTCCGAATTGGGGTAGAACATACAT
901	rapei Val Asa6iy Lys Ar of a let for AspAsp6iu LeuThr Phe LeuTyr AsaThr Phe BialiePhe AlaProSer 6ia Phe LeuProThrTrpVal Lys AsplieLeuSerVal AspTyrThr AsplieHetLys I i e LeuSer CAGTAAAC8BAAAACCCGTACGTCAGATCACCGTAGTGATGAGACTCACCTTCTTGATAACACTTTTCAAATATTTGCTCCCTTCGATTCCTACCTGGGTCAAAGACATCCTACCGGTCAAAGACATCCTTGGTCAAGACATTCTTT
1051	LysSer Jie6juLysNet6iaSer AspThr6ia6juAlaAsaAspJieValThrLeuAlaAsaLeu6iaTyrAsa6iySerThrPr0AlaAspAlaPhe6juThrLysValThrAsaJieJieAspArgLeuAsaAsaAsa6iyJieHis CCAAAAGTATTBAAAAAATBCAATCTBATACCCAAGAGECAAACBACATTBTGACCCTBGCAAATTTGCAATATAATBBCAGTACACCTGCAGGTACCAATACAAAAGTCACCAAACATTATCBACAABACTBAACAATAATBBCATTC
D1 1201	$VA_{1\overline{1}eAsaAsaLysValAlaCysGlaLeu1leNetArgGlyLeuSerGlyGluTyrLysPheLeuArgTyrThrArgNisArgNisLeuAsaNetThrValAlaGluLeuPheLeuAsplieNisAla1leTyrGluGluGlaGlaGlySer ATATCAATAACAAGGTCGCATGCCAATTAATTATGAGAGGGTCTATCTGGCGAATATAAATTTTTACGCTACACGCGCACTCTAAATATGACAGTCGCTGGTACTGTTCTTAGATATCCATGCTATTTATGAAGAACAACAGGGAA$
1351	spil Ar g Asa Ser Lys Pro Asa Tyr Ar g Asa Leu Ser Aspôlulys Asa Asa Ser Ar g Ser Tyr Thr Asa Thr Thr Lys Pro Lys Pro Lys Pro Asa Tro Gla Lys Thr Asa Asa Ser Lys Ser Lys Thr Ala Ar g Ala Nis Asa Ya CGABAAACAGCAAACCTAATTACAGBABAAATCTGAGTGATGAGAATGATTCTCGCAGCTATACGAATACAAACCAAAGCCAAAGCTATAGCTCGGAATCCTCGAAAAAACAAATAATTCGAAATCGAAAACAGGCCAGGCCTCACAAAG
	Ser Thr Ser Asa Asa Ser ProSer Thr Asa Asa Ser I le Ser Lys Ser Thr Thr 6 luProlle6 la Leu Asa Asa Lys His Asa Leu Ais der Arg Pro6 ly Thr Tyresa Thr 1 le Ser Thr Thr Phe Thr Leu Biv6 la 6 lu Leu Thr 6 lu Ser Thr Yal Asa His Ser Asa Asa
1501	TATCCACATCTAATAACTCTCCCAGGACGAGGAACGAATCGAATCAATC
1651	GINCERFORTY ALSO AND
1801	61pPheHisPhe61pAspAspThrLysThrSerIleLysValLeuHisThrProAspIleAlaTyrAspLeuLeuSerLeuAsp61uLeuAlaAlaValAspIleThrAlaCysPheThrLysAspValLeu61uArgSerAsp61yThr CAATTTCACTTCCAGBACAACCACCAAAACATCAATAAAGBTATTGCACACTCCTAACATAGCCTATBACTTACTCAGTTGGATGGATTGGCABTAGATATCACAGCATBCTTTACCAAAAACBTCTGAGGGGCACT Pati Bandaacatcaataaacatcaataaaggacatgacaatgacatgacatgacatgacatgacatgacatgacatgacatgacatgacatgacatgacatgacatgac
1951	ValleuAlaProlleVal6laTyr6lyAspPheTyrTrpValSerLysArgTyrLeuLeuProSerAsalleSerValProThrIleAsaAsaValHisThrSer6luSerThrArgLysTyrProTyrProThelleHisArglleVal GTACTT6CACCTATCGTACAATAT66A6ACTTTTACT666TATCTAAAA6GTACTT6CTTCCATCAAATATCTCCGTACCCACCATCAA TAAT6TCCATACAA6T6AAA6TATCCTTATCCTTTCATTCA
2101	AlaWisAlaAsaAla&iaThrlieArgTyrSerLeuLysAsaAsaThrlieThrTyrPheAsa&iuSerAspValAspTrpSerSerAlalieAspTyr&iaCysProAspCysLeulie&iyLysSerThrLysWisArgWisIleLys BCBCATBCCAATBCACABACAATTCBATACTCACTTAAAAATAACACCATCACBTATTTTAACBAATGTCBACTABTCBBCTABTBCCTATTBACCAATGTCCTGATTGTTTAATCBBCAAAABCACCACAAACACTATCAAA Sali
2251	61ySer ArgleulysTyr61aAsaSerTyr61uProPhe61aTyrLeuNisThrAsp11ePhe61yProValNisAsaLeuProAsaSerAlaProSerTyrPhe11eSerPheThrAsp61uThrThrLysPheArgTrpValTyrPro 66TTCACBACTAAAATACCAAAATTCATACBAACCCTTTCAATACCTACATACTBACATATTT66TCCA6TTCACAACCAACTAAGT66CACCATCCTATTTCATCTCATTTACT6AT6A8ACAACAAAAATTCC5TT666TTTATCCA
2401	LeuNisAspArgArg6luAspSerIleLeuAspValPheThrThrIleLeuAlaPheIleLysAsb6lbPhe6lbAlaSerValLeuValIle6lbNetAspArg6lySer6luTyrThrAsbArgThrLeuNisLysPheLeu6luLys TTACACBACCBTCBCBABBACTCTATCCTCBATBTTTTACTACBATACTABCTTTGTTAABAACCABTTTCABGCCBTBTTTTACAAAATBGACCBTBBTTCTGAGTATACTAACABAACCCATAGATTCCTTBAAAAA
2551	Asb6iyileThrProCysTyrThrThrThrAlaAspSerArgAlaHis6iyYelAla6iuArgLeuAsbArgThrLeuLeuAspAspCysArgThr6ibLeu6ibCysSer6iyLeuProAsbTyrLeuTrpPheSerAlaIie6iuPhe AAT66TATAACTCCAT8CTATACAACCACA6C66ATTCCC6A6CACAT66A6TC8CT6AAC68CTAAACC8TACCTTATTA6AT8ACT8CCAT6TC8CAAT6TAACT8CAAT
2701	Ser Thr Ile Val Arg Asa Ser Leu Ala Ser Prolys Ser Lys Lys Ser Ala Arg6 la His Ala61 y Leu Ala61 y Leu Asp Ile Ser Thr Leu Leu ProPhe61 y 61 a ProVall 1 e Val Asa Asp His Asa Pro Asa Ser Lys Ile His TCTACTATT&TGA&AAATTCACTA&CTTCACTAAAAGCAAAAAATCT&CAA&AACAT&CTG&CA&GACTT&GCA&GACTTC&TACTACTTC&TCACTACCT&CATGATCACT Del y
2851	BC11 ProArg61y11ePro61yTyrA1aLeuWisProSerArgAsaSerTyr61yTyr11e11eTyrLeuProSerLeuLysLysThrVa1AspThrThrAsaTyrVa111eLeu61a61yLys61uSerArgLeuAsp61aPheAsaTyrAsp CCTCBT88CATCCCA88CTAC8CCTTACATCCBTCTCBAAACTCTTAT6BATATCACTTATCTTCCATCCTTAAABAAGACAGTAAGTAACTAACTAACTATGTTATCTTCT
3001	AlaleuThrPheAsp6luAspLeuAspArgLeuThrAlaSerTyrHisSerPheIleAlaSerAsp6luIle6la6luSerAspAspLeuAspIle6luSerAspPhe6laSerAspIle6luLeuHisPro6lu6laProArg BCACTCACTITCBATBAABACTITAAACCBTITAACTBCTTCATTBCTTCATTBCBTCAATBABATCCAABAATCCAABAATCTAACATABAATCTBACCTACBACCTACBACCATBAACTACCAACCBACATCBABAACCACBABA
3151	Asa Walleu Ser Lys AlaWal Ser ProThr Asp Ser Thr ProPro Ser Thr Wis Thr 61u Asp Ser Lys Arg Wal Ser Lys Thr Asa 11e Arg Ala Pro Arg 81u Wal Asp Pro Asa 11e Ser 81u Ser Asa 11e Leu Pro Ser Lys Lys AAT8TCCTTTCAAAA6CT6T6A6TCCAACC6ATTCCAACACCTCC6TCAACTCAACAGTATC6AAAC6T6TTTCTAAAACCAATATTC666CACCCA6A6AA6TT6ACCCCAACATATCT6AAATTTCTTCCAACAA6A6
3301	ArgSerSerThrPro&lalleSerAsalle&JuSerThr&IySer&Jy&EtHisLysLeuAsaValProLeuLeuAlaProMetSer&InSerAsaThrMis&IuSerSerHisAlaSerLysAspPheArgHisSerAsp ABATCTABCACCCCCCCAAATTTCCAATATCGAGAGTACCG&TTCGBGTGGTATGCATAAATTTAAAT&TTCCTTTACTTGCTCCCAT&TCCAATCTAACACACATGAGTCGTCGCACGCCA&TAAATCTAAAGATTTCABACACTCABAC Balli
3451	SerTyr Ser Glu As b Glu Thr As b His Thr As b Val Prolle Ser Thr Gly Gly Thr As b As b Ly s Thr Val Pro Gla lie Ser As p Gla Glu Thr Glu Lys Arg lie lie His Arg Ser Pro Ser lie As p Ala Ser Pro Pro Glu TCGTACAGTBAAAATBAGACTAATCATACAAACGTACCCAATATCCAGTACGGG TBBTACCAACAAAAACTGTTCCGCAGBATAABTGACCAABBACTGAABBACTAATCACCGBTCACCGTTCCCACCBBAA Van 1
3601	Asa Asa Ser Ser Nis Asa ile Val ProlieLys Thr ProThr Thr Val Ser 6 iu 6 in Asa Thr 6 iu 6 iu Ser 1 ie 1 ie Ala Asa Leu Pro Leu Pro Asa Leu Pro 6 iu Ser ProThr 6 iu Pre Pro Asa Pro PheLys 6 iu Leu Pro AATAATCATCBCACAATATTBTTCCTATCAAAACBCCAACTACTBTTTCTBAACABAATACCBABBAATCTATCBCTBATCTCCCCACTGCTCCABAATCTCCTCCABAACTCCCCTBACCCACTACTBAACBACACTCCCCA E
3751	ProlieAsa Ser Arg61aThr Asa Ser Ser Leu61y61y11e61yAspSer AsaA1aTyrThrThrI1eAsa Ser LysLysArgSer Leu61uAspAsa61uThr61u11eLysVa1Ser ArgAspThrTrpAsaThrLysAsaNetArg CC6ATAAATTCTC6TCAAACTAATTCCA6TTT668566TAT5655ACTCTAA56CCTATACTACTACTAACA6TAAGAAAA6ATCATTA6AABATAAT6AAACT6AAATTAA66TATCAC6A6AACACAT66AATAAT665T
3901	Ser Leuß lufr ofro Arg Ser Lys Lys Arg I le Wis Leu I le Ala Ala Val Lys Ala Val Lys Ser I le Lys fro I le Arg Thr Thr Leu Arg Tyr Asp Giu Ala I le Thr Tyr Asa Lys Asp I le Lys Giu Lys Giu Lys Tyr I le Giu AGT TTAGAACCTCCGAGAATCGAAGAACGAATTCACCTGATTGCAGCGAGTAAGAGCAGTGAGGCAATCCGAGGGCAATCACCTATAATAAGGAAGAAGAAGAAGAAGAAATTATCGAG Erro I Punti Punti
4051	Alatyr Hislys GluValAsa GlaLeulealys Netlys Thr Tr pAsp Thr Asp BluTyr Tyr Asp Ar glys GluI le Asp Prolys Ar yVall le Asp Ser Net Phe Ile Phe Asp Lys Lys Ar qAsp Gly Thr Nis Lys Ar q Ar p Phe BCATACCACAABAABAGCAATCAACTGTTGAABATGATBAGAACTTGBBACACTGACBAATATTATGACABAAAAABAAGCATAABAAGAGTAATAAACTCAATGTTATCTTCAACAABAAAGGTBACBGTACTCATAAACBTABATT
4201 T	Val AlaArg6iyAsplie6jaHisProAspThrTyrAspSer6jyNet6iaSerAsaThrValHisHisTyrAlaLeuNetThrSerLeuSerLeuAlaLeuAspAsaAsaTyrTyrlieThr6iaLeuAsplieSerSerAlaTyrLeu GTTGCAABABBCGATATTCAGCATCTGACACTTACGACTCABBCATGCCAATACCBTACATCACTATGCATTAGACATACTCGTCACTTGCACTAGACATAACTACTATTACACAATAABACATATCTTCCBGCATATTB Sabi
4351	Tyr AlaAsplieiysêluêluleuTyr lie Ar gêr ofro Nisleuêly Net Aso Asplysieulie Ar gleulysiysêr leuTyrêluleulysêlo Ser êly AlaAso Trp TyrêluThr lielysêr Tyrleulieêlo êlo TATBCABACATCAAABAACATTATACATAABACCTCCACCACATTABBAATBAAT
4501	Cys8iyWet8iu8iuValArg8iyTrpSerCysValPbelysAsbSer6ibValThr1ieCysLeuPheValAspAspNetValLeuPheSerLysAsbLeuAsbSerAsbLysArg1ie1ie8iuLysLeuLysNet8ibTyrAspThr T6568TA168AAAABTTC8F68AT66TCAT6CBTATTTAAAAATABTCAAGTBACAATTT6TTATTC9TA6AT8ATA158TAT6TTTABCAAAAATCTAAACAAAABAATTATA6AAAACCTTAABAT6CAATAC8ACACC Hind111

FIG. 1. (Figure continues on the next page.)

	Lys I ie I ie Asa Leu 6 i y 6 iu Ser Asp 6 iu 6 iu 1 ie 6 ia Tyr Asp I ie Leu 6 i y Leu 6 iu 1 ie Lys Tyr 6 ia Arg 6 i y Lys Tyr i	NetLysLeu6iyNet6iuAsaSerLeuThr6iuLysIieProLysLeuAsaVaiProLeuAsaPro
4651	51 AÁBATTATANATCTAGSÓGAAAGTGATBABGANATTCAATÁTGAĆATACTTBSÓTTABANATCAÁATÁTCAAABÁBGTAÁATÁC	ATBAÀATTABETATBBAAAACTCATTAACTBABAÀAATACCCAÀATTAAACBTACCTTTBAATCCA

Lys Ely Arg Lys Leu Ser Ala Pro Ely Eu Fro Ely Leu Tyr II e AspEla AspEla Leu El al I e AspEla Lys Ela Lys Ela Lys Ela Lys Leu II e Ely Leu Ala Ser Tyr Val Ely Tyr Lys Phe Arg Phe 4801 MAADBAABAAAACTTAGCGCTCCAGGTCAACCAGGTCTTTATATAGACCAGGATGAACTAGAATAGATBAAGATGCAAAGGGAGGAGGTACATGAAGTGAATGGATG

AsstysteuvalAlalieSer AspAlaSer Tyr BiyAssBisProtyr Tyr LysSer BislieBiyAss IIePbeleuLeuAssBiyLysValIIeBiyBiyLysSer ThrLysAlaSer LeuThrCysThrSer ThrThrBieLalaBiu 5101 AATAAACTABTCBCAATAABCBATBCTTCATATBGTAACCAACTATTACAABTCACAAATTBBTAACATTTTCCTACTCAACBBAAAABTGATTBBABGAABGCCAAABBCCTTCBTTAACATBCCAACTACAABTCAACAATTBBTAACATTTTCCTACTCAACBBAAAABTGATTBBABGAAABTCBACAAABBCCTTCBTTAACATBCCAACTACAACTACAABTCAAABTCAACAATTBBTAACATTTTCCTACTCAACBBAAAABTGATTBBABGAAABTCBACAAABBCCTTCBTTAACATBCAACTTCCAACTACABAABCABAA 5101 AATAAACTABTCBCCAATAABCCBATBCTTCATATBGTAACCAACTTACAABTCAACAATTBBTAACATTTTCCTACTCAACBBAAAABTGATTBBABGAAABTCBACAAABBCCTTCBTTAACATBCCAACTACCAACAABTCAAABTCAACAATTBBTAACATTTTCCTACTCAACBBAAAABTGATTBBABAAABTGATCBACAAABTCBACAAABBCCTTCBTTAACATBCCAACTACCAACTACAABTCAACAATTBBTAACATTTTCCTACTCAACBBAAAABTGATTBBABGAAABGTGACCAAABBCCTCBATABCCAACTTCCAACTBCAACAABTCACAAABTCAACAATTBBTAACATTT

- Sali Hpai IleNisAlaValSer6luAlalleProLeuLeuAsnAsnLeuSerNisLeuVal6ln6luLeuAsnLysLysProlleIleLys6lyLeuLeuThrAspSerArgSerThrIleSerIleIleLysSerThrAsn6lu6luLysPheArg 5251 ATACAC6CAGTCABTGAABCTATACC6CTATTGAATAACCTCAGTCACCATGTGAGAAACCCAATTATTAAAGCCTACTGATAGTABATCAACGATCABTAGAATTAA
- Asa ArgPhePheBiyThrLysAlaNetArgLeuArgAspBiuValSerBiyAsaAsaLeuTyrValTyrTyrlleBiuThrLysLysAsaIleAlaAspValNetThrLysProLeuProlleLysArgPheLysLeuLeuThrAsaLys 5401 AACABATTITTEBCACAAABBCCATEABACTATEABAGTATCABETAATAATATTATACETATACTACATCBABACCAABAABAACATTECTBATEABAACCAATBABACTATCABBAABATTTAAACTATTAACTAAT

- 5851 TACCTAATATTATAGCCTTTATCAACAATGGAATCCCAACAATTATCTCAACATTCACCCATTTCTCA

FIG. 1. Sequence of Ty912. The coding strand of the element is shown. The *delta* terminal repeats are underlined. The *tya912* and *tyb912* genes are translated above their DNA sequences. Restriction sites appear below the DNA sequence. The sequence of over 93% of each strand has been directly determined. The regions of peptide homology referred to in the text are indicated above the amino acid sequence and marked to identify them: DNA for DNA-binding protein homology (14); M3, M4, R2, I, II, and III for reverse transcriptase homology (15).

region upstream of the AUG is translated, suggesting that the AUG is not the site of initiation of tyb912. From inspection of the DNA sequence we could not assign a site for initiation. The tyb912 open reading frame could encode a protein of as much as 151,200 daltons. There are no other reading frames longer than 250 bp in either strand of the element.

The tya912 Protein Includes Homology to DNA Binding Proteins. The tya912 protein includes two regions of unusual primary sequence. A region of 125 amino acids, beginning at amino acid 58, is 21.6% proline. The carboxyl-terminal region of 90 amino acids, composed of 20.2% basic residues, is reminiscent of a DNA-binding protein. Upstream of this basic region, from residue 313 to residue 327 (Fig. 1), is a sequence similar to the consensus for DNA-binding proteins from bacteria (14). The homology is found within a region predicted by the Chou and Fasman method (17) to include two potential α -helical regions separated by a β -turn, with the homology occurring at the correct position with respect to these putative secondary structural features. Of the 10 amino acids that have been shown or predicted to interact with DNA (14, 18, 19) 9 occur within the putative two α helices of the tya912 protein (Fig. 1). The primary and secondary structural features suggest that the protein may be a DNA-binding protein.

The tyb912 Protein Shows Homology to Retroviral Reverse Transcriptase. Given the similarity between Ty elements and retroviruses, we hypothesized that the tyb912 protein might be homologous to retroviral reverse transcriptase. Toh *et al.* (20) and Patarca and Haseltine (15) have found scattered primary sequence similarities among both retroviral reverse transcriptases and DNA polymerases of double-stranded DNA viruses. We find that the region from amino acid 835 to amino acid 1046 of tyb912 includes sequences similar to four homology regions referred to as R2, I, II, and III (15) and two additional regions (M3 and M4; ref. 15) from amino acid 306 to amino acid 359 that we find are similar to reverse transcriptases but not DNA polymerases (see Fig. 1). This suggests that the product of the tyb912 gene may be a reverse transcriptase.

An Unusual Mode of Expression for the tyb912 Gene. The tyb912 gene is unusual in two respects: first, the 5' end of the tyb912 open reading frame overlaps the 3' end of the tya912 open reading frame for a distance of 38 bp; second, the first in-frame ATG codon occurs 921 bp from the beginning of the open reading frame. It is doubtful that this is the initiator codon for the tyb912 gene since it is known that eukaryotes

normally initiate translation within 40-80 nt of the 5' end of the transcript (16).

To examine the expression of the *tyb912* gene we constructed a gene fusion with the *lacZ* gene of *Escherichia coli*. In this construction, pPF350, the *lacZ* gene is fused to the *tyb912* coding frame at a *Bgl* II site (nt 1702), upstream of the initial ATG codon but downstream from the 3' end of the *tya912* gene (Fig. 1). When pPF350 is used to transform yeast, significant levels of β -galactosidase (1004 units) are detected, showing that the *tyb912* coding frame is expressed and that translation of the gene product must initiate upstream of the first in-frame ATG codon. We have sequenced the fusion junction of pPF350 to confirm that *lacZ* expression is not a cloning artifact.

The simplest explanation for the expression of the tyb912 coding frame would be that the gene is spliced to an upstream initiator, most likely that used by the tya912 gene. To investigate this possibility S1 nuclease mapping was performed, using the tyb912::lacZ gene fusion of pPF350 (Fig. 2). DNA·RNA hybrids protected from S1 nuclease degradation were analyzed in neutral and alkaline agarose gels followed by Southern blotting. A lacZ-specific Southern probe reveals a single band in both gels of the size expected for an unspliced transcript. When a Southern probe specific for the 5' end of Ty912 was used, this band, plus another, was observed in both gels. The second band is due to protection of the S1 nuclease probe by wild-type Ty mRNA. No other protected species were observed. S1 nuclease analysis was also performed by using a 5'-end-labeled restriction fragment extending from the Sph I site in the tya912 gene to the Pvu II site in lacZ (data not shown). The sequence ladder of this portion of DNA was used as size markers for the protected hybrids. No evidence of spliced RNA was observed.

It appears from these experiments that splicing of the tya912 and tyb912 genes does not occur. Nevertheless, we have shown that tyb912 expression is dependent upon translation of the tya912 gene by constructing a frameshift within this open reading frame. This alteration in the tya912 reading frame of pPF350 introduces an in-frame TAA stop codon 4 bp downstream of the *Sph* I site at nt 1218 (plasmid pJC8). Virtually no β -galactosidase activity (<1 unit) is observed when this plasmid is present in yeast. Thus tyb912 gene expression is eliminated by early termination of the tya912 gene and cannot be due to translation initiating upstream of the *Bgl* II site within the tyb912 open reading frame itself (e.g., at a codon other than ATG).



One interpretation of this result is that the primary tyb912 gene product is a fusion protein that is produced by translation continuing from the tya912 coding frame into the tyb912 gene. To test this idea we have studied the Tyb- β -galactosidase fusion proteins produced in strains carrying pPF350. After purification by affinity chromatography and blotting analysis, two major bands, whose molecular weights are estimated to be 175,000 (±6000) and 123,000 (±2000), and several minor bands of intermediate sizes were observed (Fig. 3). No additional polypeptides were found to cross-react in extracts not affinity purified. A protein produced by initiating at the first ATG of the tya912 coding frame and continuing into the tyb912::lacZ gene would have a predicted molecular weight of 172,000. Our observed value (175,000) is in good agreement with this, taking into account the limits of resolution for proteins this large on NaDodSO₄/polyacrylamide gels. We therefore propose that the tyb912 gene product is expressed by translation initiating at the beginning of the tya912 coding frame and continuing into the tyb912 gene by frameshifting in the region of overlap between the two genes.

The second major band has a molecular weight of 123,000, 3000 larger than expected for β -galactosidase from this vector. A protein of this size would occur if the largest fusion protein was processed by proteolytic cleavage near the carboxyl end of the *tya*-encoded portion. The minor bands of intermediate sizes presumably represent multiple specific cleavages within the amino terminus of the fusion protein, since no polypeptides smaller than β -galactosidase are observed.

DISCUSSION

The nucleotide sequence presented here reveals the unusual genetic organization of the yeast transposon Ty912. We have identified two overlapping reading frames, tya912 and tyb912, a structure similar to that found in the retrovirus

FIG. 2. S1 nuclease analysis of the tyb-912::lacZ fusion transcripts. The fragments protected from S1 nuclease degradation by a Sac I restriction fragment from pPF350 were separated in duplicate neutral (A) and alkaline (B) gels and then analyzed by Southern blotting. Each pair of filters was then hybridized to either a lacZ-specific probe (I and III) or a probe specific for the 5' end of Ty912 (II and IV). Lanes 2, 4, 6, and 8 show fragments protected from S1 nuclease (750 units/ml). For comparison the Sac I restriction fragment was run on each gel (lanes 1, 3, 5, and 7). The sizes given (in kb) were estimated by using HindIII-digested DNA of bacteriophage λ as markers. A restriction map of the relevant portion of the fusion plasmid pPF350 is also shown. The sizes of S1-nuclease-protected fragments expected from unspliced hybrid gene transcripts (X) and unspliced wild-type Ty mRNA (Y) are indicated.

Rous sarcoma virus (RSV) (21). The analogy with metazoan retroviruses is further strengthened by our demonstration of a tya::tyb912 fusion protein, which is reminiscent of the retroviral gag-pol proteins. The predicted tya912 protein exhibits several structural features characteristic of DNA-bind-



FIG. 3. Blot analysis of tyb912::lacZ fusion proteins. Hybrid proteins, purified by affinity chromatography, were electrophoresed in NaDodSO₄/9% polyacrylamide gels and blotted to nitrocellulose. The proteins on the filters were allowed to react with mouse IgG anti- β -galactosidase and the bands were visualized by using goat anti-mouse IgG conjugated with horseradish peroxidase and Bio-Rad color development reagent. Lane 1, proteins isolated from a yeast transformant of pPF350; lane 2, high molecular weight markers, including β -galactosidase (β GZ; molecular weight, 116,000).

ing proteins. Examination of the tyb912 sequence reveals scattered homology to several reverse transcriptases and DNA polymerases. The degree of similarity found is comparable to that displayed by any of these enzymes when compared to the rest of the group, suggesting that this observation is significant. Thus sequence analysis of the Ty912-encoded proteins suggest a possible role in the transposition of the element itself.

We have also demonstrated an unusual mode of expression for the tyb912 gene. We have shown that the product of this gene is a tya912::tyb912 fusion protein, yet we observed no evidence of spliced RNA. It is possible that the protein results from the excision of an intron too short to be detected by S1 endonuclease. However, we consider this unlikely due to the high sensitivity of the technique (9) and because of the absence of sites known to be required for splicing of yeast nuclear mRNA (i.e., donor, acceptor, and T-A-C-T-A-A-C sites). Furthermore, the efficiency of mRNA splicing in yeast appears to be greatly reduced with a diminishing size of intron (22). Our alternative explanation is that the tya-912::tyb912 fusion protein is produced by translational frameshifting within the region of overlap between the two genes.

Several other examples of translational frameshifting are known. The lysis genes of the RNA phages f2 and MS2 (23, 24) are expressed by frameshifting within the corresponding upstream coat gene. Spontaneous readthrough of leaky frameshift mutations occurs within the oxil gene of yeast mitochondria (25). The addition of a carboxyl-terminal extension onto the phage T7 major coat protein also occurs by a frameshift during translation (26). The level of frameshifting observed in each case is due to translational misincorporation (27).

Another system in which translational frameshifting allows the expression of alternative proteins is the gag-pol genes of RSV (21). The primary translation product of pol is gag-pol fusion protein. The expression of the fusion is usually attributed to a splicing event, although there is no direct proof of this (21). The RSV genome includes a 58-bp overlap between the gag and pol genes similar to that found in Ty912. Thus it is also possible that expression of the *pol* gene occurs by frameshifting within the overlap, as was proposed by Schwartz et al. (28). A model of translational misincorporation may also account for the expression of the *pol* gene of Moloney murine leukemia virus. Here the gag and pol genes are in the same frame but separated by a single UAG nonsense codon. Translation of the latter gene may occur by readthrough of this codon. It is possible that mechanisms involving translational infidelity have a general role in the expression of proteins required in low amounts.

We have also sequenced a Ty2 element, Ty917, and found that it too contains overlapping reading frames (unpublished data). When this sequence is compared to that of Ty912 the homology found in the overlap is no greater than that of the adjacent 800-bp region. We therefore believe that the primary structure of the overlap region is not important in promoting frameshifting. Rather, our model involves translational misincorporation at the UGA terminator codon of the tya912 gene. Weiss (29) has described a general mechanism to explain how certain normal tRNA molecules are able to promote ribosomal frameshifting. This is based upon studies that identify these so-called "shifty" tRNAs, and their corresponding "shifty" codons (30, 31). The conserved U33 nucleotide immediately 5' to the normal anticodon in these tRNAs is proposed to base pair with the "shifty" codon. By applying the "reciprocating ratchet" model, in which translocation is caused by an allosteric transition of the anticodon loop (32), Weiss has predicted that this offset anticodon-codon pairing would cause a -1 or +1 frameshift. The direction of the frameshift produced is determined by which of two possible allosteric conformations is energetically more favorable, and this in turn is dependent on the particular tRNA and codon involved.

It is possible that tyb912 gene expression occurs by such a mechanism. Binding of a leucyl-, methionyl-, or valyl-tRNA, each of which contains an offset 3'-ACU-5' anticodon, to the 5'-UGA-3' terminator codon of the tya912 gene could occur. This interaction would presumably be in competition with the binding of termination factor to the ribosome such that translation of tyb912 is dependent on the relative binding rates of tRNA and factor. Further experiments are necessary to define the actual mechanism of tyb912 gene expression.

We thank Lee Scionti and Celeste Sticca for their capable technical assistance and Janice Seagren for typing the manuscript. This work was supported by Grant GM 29480 from the Public Health Service and Grant 6-30151 from the American Heart Association.

- Cameron, J. R., Loh, E. Y. & Davis, R. W. (1979) Cell 16. 1. 739–751.
- 2. Roeder, G. S. & Fink, G. R. (1980) Cell 21, 239-249.
- Farabaugh, P. J. & Fink, G. R. (1980) Nature (London) 286, 3. 352-356.
- 4. Gafner, J. & Philippsen, P. (1980) Nature (London) 286, 414-418.
- 5. Elder, R. T., Loh, E. Y. & Davis, R. W. (1983) Proc. Natl. Acad. Sci. USA 80, 2432-2436.
- 6. Varmus, H. E. (1983) in Mobile Genetic Elements, ed. Shapiro, J. A. (Academic, New York), pp. 411-503
- 7. Guarente, L., Yocum, R. & Gifford, P. (1982) Proc. Natl. Acad. Sci. USA 79, 7410-7414.
- 8. Maxam, A. & Gilbert, W. (1980) Methods Enzymol. 65, 499-560
- 9 Favaloro, J., Treisman, R. & Kamen, R. (1980) Methods Enzymol. 65, 718-749.
- 10. Germino, J., Gray, J. C., Charbonneau, H., Vanaman, T. & Bastia, D. (1983) Proc. Natl. Acad. Sci. USA 80, 6848-6852.
- 11. Towbin, H., Staehelin, T. & Gordon, J. (1979) Proc. Natl. Acad. Sci. USA 76, 4350-4354.
- 12 Miller, J. (1972) Experiments in Molecular Genetics (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 352-355.
- 13. Donahue, T. F., Farabaugh, P. J. & Fink, G. R. (1982) Gene 18, 47-59.
- 14. Pabo, C. O. & Sauer, R. T. (1984) Annu. Rev. Biochem. 53, 293-321.
- 15. Patarca, R. & Haseltine, W. A. (1984) Nature (London) 309, 728.
- Kozak, M. (1983) Microbiol. Rev. 47, 1-45. 16.
- 17. Chou, P. H. & Fasman, G. D. (1978) Adv. Enzymol. Relat. Areas Mol. Biol. 47, 45-148.
- Ohlendorf, D. H., Anderson, W. F., Fischer, R. G., Takeda, 18. Y. & Matthews, B. W. (1982) Nature (London) 298, 718-723.
- 19. Weber, I. T. & Steitz, T. A. (1984) Proc. Natl. Acad. Sci. USA 81, 3973-3977
- 20. Toh, H., Hayashida, H. & Miyata, T. (1983) Nature (London) 305, 827-829.
- 21. Weiss, S. R., Hackett, P. B., Oppermann, H., Ullrich, A., Levintow, L. & Bishop, J. M. (1978) Cell 15, 607-614.
- 22. Langford, C. J., Klinz, F. J., Donath, C. & Gallwitz, D. (1984) Cell 36, 645-653
- 23 Berem, M. N. & Blumenthal, T. (1979) Cell 18, 257-266.
- 24. Kastelein, R. A., Remant, E., Fiers, W. & van Diun, J. (1982) Nature (London) 295, 35-41.
- 25. Fox, T. D. & Weiss-Brummer, B. (1980) Nature (London) 288, 60-63.
- Dunn, J. J. & Studier, F. W. (1983) J. Mol. Biol. 166, 477-535. 26.
- Roth, J. R. (1981) Cell 24, 601-602. 27.
- Schwartz, D. E., Tizard, R. & Gilbert, W. (1983) Cell 32, 853-28. 869
- 29. Weiss, R. B. (1984) Proc. Natl. Acad. Sci. USA 81, 5797-5801.
- Atkins, J. F., Gesteland, R. F., Reid, B. R. & Anderson, 30. C. W. (1979) Cell 18, 1119-1131.
- 31. Weiss, R. & Gallant, J. (1983) Nature (London) 302, 389-393.
- 32. Woese, C. (1970) Nature (London) 226, 817-820.