Nucleotide sequence characterization of Ty 1-17, a class II transposon from yeast

John R. Warmington, Richard B. Waring, Carol S. Newlon⁺, Keith J. Indge and Stephen G. Oliver*

Department of Biochemistry and Applied Molecular Biology, University of Manchester Institute for Science and Technology, P.O. Box 88, Manchester M60 1QD, UK, and ⁺Department of Microbiology, UMDNJ-New Jersey Medical School, 100 Bergen Street, Newark, NJ 07103, USA

Received 29 July 1985; Revised and Accepted 3 September 1985

ABSTRACT

We have determined the nucleotide sequence of a class II yeast transposon (Ty 1-17) which is found just centromere-distal to the LEU2 structural gene on chromosome III of Saccharomyces cerevisiae. The complete element is 5961 bp long and is bounded by two identical, directly repeated, delta sequences of 332 bp each. The sequence organization indicates that Ty 1-17 is a retrotransposon, like the class I elements characterized previously. It contains two long open reading-frames, TyA (439 amino acids) and TyB (1349 amino acids). In this paper, the sequences of the two classes of yeast transposon are compared with one another and with analogous elements, such as retroviral proviruses, cauliflower mosaic virus and copia sequences. Features of the Ty 1-17 sequence which may be important to its mechanism of transposition and its genetic action are discussed.

INTRODUCTION

Mobile genetic elements within the nuclear genome of eukaryotes were first recognized in a higher plant, maize, by McClintock (1,2). Since those original studies such elements have been discovered throughout the eukaryotic evolutionary scale from yeast to Drosophila and man (3). A number of these elements, including the retroviral proviruses (4), cauliflower mosaic virus (5), copia-like elements in Drosophila (6) and the Ty elements of yeast (7) show an overall similarity in DNA structure and, indeed, are somewhat reminiscent of prokaryotic transposons (8). While it is well established that both retroviruses (4) and cauliflower mosaic virus (5) employ RNA as an intermediate in their replication and transposition, it was not clear that non-viral elements such as copia and Ty were capable of reverse transcription (9). Recently, however, Boeke et al. (10) provided a direct demonstration that a class I yeast Ty element (TyH3) transposes via an RNA intermediate and coined the term 'retrotransposon' to describe this general class of non-viral mobile genetic elements.

The Ty elements of the yeast Saccharomyces cerevisiae are present at an average copy number of ca. 30 per haploid genome in most strains (11). A generalized Ty element can be described as a 5.2 kb unique 'epsilon' region flanked by identical direct repeats, called 'delta' sequences, of ca. 340 bp (11). However, the family of Ty elements shows a

sequence heterogeneity which is exhibited in their restriction maps (Fig. 1), such that they can be divided into two broad types of element, class I and class II (12).

Heteroduplex analysis has shown that these two types of element show large regions of non-homology. This apparent heterogeneity between the two types of element may be misleading, however.

In a previous publication (13), we compared the first open reading-frame (ORF) of a typical class I element, Ty 1-15 (ref. 12) with that of a typical class II element, Ty 1-17. We found that these two ORFs, designated TYA15 and TYA17, showed considerable sequence divergence at the DNA level. However they each encoded p1 proteins of ca. 50 kd in molecular weight which showed considerable amino acid sequence homology and a high degree of structural similarity. Such conservation implies that the p1 proteins play some important role in Ty biology and a nucleic acid binding activity has been demonstrated for them (see Discussion). The availability of the complete nucleotide sequence for two class I elements, Ty 912 (ref. 14) and Ty 109 (ref. 15) suggested that further sequence comparisons between the two types of element might shed more light on the functional importance of the different regions of the yeast transposon. To this end, we present here the complete nucleotide sequence of Ty 1-17 and compare this class II element both with the class I yeast transposons and with the sequences of other eukaryotic mobile genetic elements of the same general class.

MATERIALS AND METHODS

DNA Sequencing

The sequence of the Ty 1-17 transposon was obtained using the random sequencing approach of Deininger (16). The transposon was contained on two BamHI fragments of chromosome III, A5C and G4B, which had been cloned into the Ylp5 yeast-E. coli shuttle vector (17). The uniqueness of the BamHI site at the junction between these two fragments was confirmed by examination of an EcoRI bank of the ring form of chromosome III cloned into the bacteriophage λ vector, charon 4A. The two BamHI fragments were purified from agarose gels, circularized and broken into random fragments by sonication (16). These fragments were size fractionated by agarose gel electrophoresis and suitably sized fragments (300-1000 bp) cloned into the SmaI site of M13 mp8 (18) by blunt end ligation. The inserted DNA fragments were then sequenced by a modification (19) of the dideoxynucleotide chain termination method of Sanger et al. (20). Assembly and analysis of the sequence data was performed on a Prime 9950 computer as described in ref. 13. Each nucleotide was sequenced at least twice and greater than 95% of the sequence was determined in both orientations.

Chemicals and enzymes.

Restriction endonucleases were obtained from BCL Ltd., Pharmacia and BRL Ltd.

T4 DNA polymerase was purchased from Pharmacia. T4 DNA ligase and the Klenow fragment of DNA polymerase I were from BCL Ltd. [$\alpha^{32}\text{P}$] dATP (sp. activity 400 Ci mMol^{-1}) was purchased from Amersham International.

RESULTS

Location of the Ty 1-17 transposon

The class II transposon Ty 1-17 is located on chromosome III of *Saccharomyces cerevisiae*. The source DNA for our sequence of this element was a circular derivative of chromosome III formed by an intrachromosomal recombination event between HML, on the left arm, and MAT, on the right arm, of the chromosome (21). This 63 μ circular derivative represents a unique molecular species within the yeast nucleus; this has been purified and used to construct gene banks of this portion of chromosome III (17). Ty 1-17 (12) is located on the left arm of chromosome III ca. 8 centimorgans from the centromere. It is flanked on the centromere proximal side by a tRNA_{3^{leu}} gene and the structural gene LEU2 (22 and our own unpublished observations). On the centromere

Figure 1a. Generalized diagram of a Ty element with representations of its single large transcript and the two overlapping open reading-frames, TYA and TYB. **b.** Restriction site maps of the class II Ty element, Ty1-17 and the class I element, Ty 109. A schematic diagram of the heteroduplex formed by annealing single-stranded DNA molecules from each class of element (12) is shown between the two maps B=BamHI; Bg=BglII; C=ClaI; E=EcoRI; H=HindIII; K=KpnI; P=PstI; Pv=PvuII; S=SalI; Sa=SacI; X=XhoI; Xb=XbaI.

TGTTGGATTAAMAACTAACTATCTACTACAGTACTATTACGTTACTAGTATATATATCATATACGGCTGTAGACAGATGCCAAMATGACAGAAATGATCATCAATATATCGCAAGC	120
TCAAACCCAGGAGTCATAACTCAATAGGATCAATGATATATTAACATATAAAATGATGATATATATATATATACAGTTCTGACAGATTCCGATCCCTTTATGCGATTCTAAATCTCTC	240
MESQDLHQNPRSLIHGSAYASVTS	23 (AS)
CAGCAGACCTTCTACTATATCTACATACCTAAATATATTCGCTTATAAAAATGGAATCCAGCAATACATCAAAATCCAGCTCTCTCTCATGCTAGCCGATGCTCTGGTATCTCT	360 (MS)
REVP SNDDP L A V S A S N L P E F D R D S T K U N S O Q E T T P G T S A V	43 (AS)
AGGCACTCCCATCAATCAAGATCCGTTAGCCCTCTCAGCTTCCAAATTTAGCCGAATTTGATAGAGATTCCACTAGGTTATTTCTCAACAGACAGACACACTCGGCATCAGCTGTT	480 (MS)
PENHHVVSPQAPASVPPPPNQYQHQHNMPTAMKASNAH	103 (AS)
CAGACAGCATCATCTGCTCTCTCTCAAGCTCTCAGCTACAGATCCAGACAGACAGGCGATGATGACCCCAAAAGAGCTATCGCCCTTAACTGGCAAAH	400 (MS)
YQPSMHTCSHYQTSAPAYYQDPHYPLQYIPLPSTSSPD	143 (AS)
TACCACAAACCCCTCTATGACAGCTGTTCACTATTCAAGGCTACCTGGTATATATACAGGAGCCACACTACCTTCCGACAGTATATCCACCACTCAGCTATCTCTCACTCAT	720 (MS)
PIDSONQHSEVPQAEKVRNVLPPHPTLTSDENFSTHUKF	183 (AS)
CCATCGATTCCAGAGTCAACACTTGAAGTACTTCAAGCTAGACAGAGGSEAGAAATAGCTTACCCACACACTTTTACATCAGATGAAACTTTTCTACATCGGTAATTT	840 (MS)
YIRFLKNSNLGQIIPNDGCEIKRKNQYEEHAAITNFOAF	923 (AS)
TACATCAGATTTTTCAGACAGCTTATCTCGCTGACATCTCCAAATCAGGCGTAAATACAGACAAATGACTATTAGACAGATCGGCTATATATCAATCTATCTCGACATTT	260 (MS)
APFHLLFTMWKQILEINYADILTVLCKSVSKMOTNQELK	263 (AS)
GGCCATTTTATTTATTCGCACATCGCTAAACAAATTTAGAAATTTATTTATCTGACATCTTACAGCTCTTTTAAAGATCTGCTCAAAATCGCAACTAACATCAGAAATTAAG	1380 (MS)
DMIALAHLELYGCGSTASATDFEITVSTIQRLKENNINAVSDR	303 (AS)
GATTCGATAGCTTCTCCACACTTCCAGTACAGGAGGACTGCTCTGATACATTTGAAATACAGTACAGTACATCTTCACAGCTTAAGAGKAAACCAATATATCTTAGGACAG	1200 (MS)
LACQLILGLSGDEFKYLRHOYRKTNNKLSLQFAEIQIY	343 (AS)
TTGCGCTTCAACTTACTTAAGGCTATCTCGGTCTTCAATACCTAGCTATTAATATCTGCAACAAACAGCATGAATCTTCCCATTTATCTCGAATCTGATATATAT	1920 (MS)
DENKXIMNHNKPSQYKQHSEYKKNVSRTPNTTNTKVTTRY	383 (AS)
GAGCAATTAATATCTGATCTTAATTAAGCCCTCCCATACAGACAGACAGGATCAAAATGTTTTCTGCACTCTCCAAACAGACTAGACAGAGGCTTACAGCTCTGAATAT	1440 (MS)
QRTNSSKFRAAKAHNIATSSKFSRVNNDHINESTVSSOYL	423 (AS)
CAGAGAACTCTCAAAACAGACAGCAAAACTCACAATTTCTACATCTACTAATTTCTCAGGCTGACAAATGATGATACATTAATGATCAAGCTTTATCACAATCTTA	1560 (MS)
SDDNELSLRPATERI*	439 (AS)
AAITNLVLGLGQOQKAGASATITAGCCACACACACACATAGCTGATCTGATCACTTCTATTGATTTAGGACGATCTCGCAAC	400 (AS)
AGCCATACAGCAACTTCTTACGACAGACAGACACATITAGCCACACACACACATAGCTGATCTGATCACTTCTATTGATTTAGGACGATCTCGCAAC	1680 (MS)
LURSAHYLHHAATPNSIINIUDAKQNDIPINAGNLHFNFO	800 (AS)
GCTTCTGATGACGCCATTTATTCAGCATCGAACGCCATTTGAAATTAATACTACTGCTCTCAACAAACAGCATCTCTTAATATGCCATCTGTAATCTTCACTTCACTTTCA	1800 (MS)
WGTXTSINALHTPNIAFYDLISLSELANDHITACFTRTNLE	1200 (AS)
CAGGCGACCAAGACATGATTAAGAGCATACACACACACACATGATCTGATCTTAATTAATTTGAGTGAGCTGCTGATCAAAATATACAGCTGCTTTACACAGACATCTTGA	1920 (MS)
RSDDCTVLAPLIVXKNGDFYFWSLKSLIPSHISLKTINNUNXS	1600 (AS)
AGATGAGCTGTACAGCTCTGCTCCCATCTAGACATGACGCTTTTACTGTTATCAAAATACCTTATTTCTGCGCATTTCAAGGTAACAAATAGCAAGCTCGCAAAAG	2400 (MS)
KSUNKYPYPLIHRMLGHAMFRSTONSLKNNAVYTLKESDI	2000 (AS)
CAAAAGCTTAATATATCATATCCCTTAATACATCCATGCTCTGGACACTTACCTCCGAGATATCAGAGCTCTTACAGACATCAGCTTACATATTTGAGGCAATGCGATAT	2160 (MS)
EMSNACTATYQCPCDLICKSTKRRHUNXGSRILKYQESYEPFOY	2400 (AS)
TEAGTCTCTAACGACCATATATGCTCTGACTCTTAATCGCAAAAGCACACACATGACATCTCAAGGATCAGCATTAAGTACAGCAATCATAGCTTTTTCAGTA	2280 (MS)
LHTDIFGPHVHLHPKSAPSYFISFTDENTRFQWVYPLNDR	2880 (AS)
CTTCATACGATATATTTGCTGCTTACATCTTACCCAAAATGACACTTCTTACTTTATTTGCTTTAGAGATGAGAAACAGATCTCAATGGCTACCCATTTACGAGCTGCG	2400 (MS)
ESILIHVFSTILAFIKNOFMARVLVIONDRGSEYTNXYTLH	3200 (AS)
TCAGATCTTWTCTCANTTTTTTACATCGATATTAGCATTTATTAGACCAATTAATCATGCTGCTCTAGTTATCCAGATGATGCTGCTGCTCCGATCTACACAAATCTTTCA	2520 (MS)
KFFTNRGITACYTNTTADSRANGVAERLNRLLNDRCTLLH	3600 (AS)
TAGCTTTTACAGACAGCATTTATTTAGCTGATTCACACAGCCAGATTTACAGACAGCTGCTGATGACATTAATCTGACTTTTATTAAGCATTTCTGACACATCTTCA	2640 (MS)
CSGLPNHLLWFSAVEFSTIIRNLSLVSPKNDKSAARQHAGCLA	4000 (AS)
TTGCACTGCTTACCAATCATCTATGCTTCTGACAGCTGATTTTTTACTTATATCAAAATTTATTTACTTCAACAAAAACGATTAATCCGCGCAGAGACAGATCAGCTTTAGC	2760 (MS)
GIDITATCTTALPFGOPVUWVNNHNPDSILHPRGIFGYALHPSRN	4400 (AS)
TGCACTGCACTTACTTACTACTTTTGGTCAAGCCTTATTTTATTAACAGATTAATCCGCACTCGAAATACATCTCTGCGCATCTAGCTTACGCTTTACATGCTCAGAA	2880 (MS)
SYGYIYLPSLKXVDTTNNYUULQDKSLKDFNYDTLTF	4800 (AS)
CTCTATGCTATATATCTATCTTCAATATTTAAAAAGACATGATATACCATTCGTTATATTTACAGACAGCAATCAATTTGACCCATTTCAATATGATGACTACTCTT	3900 (MS)
DDPRLRLTAHNSFIEQNETEQSYDQNTESDNDYQSEIEI	5200 (AS)
TGATGATGCTCTCATCTTTTACAGCCCATTAACCATCTTTTATTTCAAAATGAGAGGAGAGCTGATGATGATCAAAATACAGATTTGATGATGATCTATCAATCCGACATGAAAT	3120 (MS)
MSDPLTNTLNTFSSQVPPPLQLDKPEVQURAPKEVDADISEY	5600 (AS)
AACTCTGCTTCTACTAGACACTTATCTCCCATTAATAGGCTTTAGCAATTTACAGACAGCAGCAGCTCGCAAAATAGCTGACCAACAGAGCAATTTGCGCAGATCTTGATTA	3240 (MS)

N I L P S P U R S R T P H I I N K E S T E M G C T V E S D T T S P R H S S T F T 609 (AAS)
 C A T A T T C T C A T T C T C T A C G A T C T C T A C A C C C A T A T T A T A A M G A C A T C C A A T G C G T A C C G T G A A T C A G A T A C T A C C T A C A G A C T C T C T A C C T T A C 3360 (NAS)

A R M Q K F P G S P M D M I D L T S G D R V N Y G L E N I K T T R L G G T E E P 640 (AAS)
 T G C A C A A A C A A A G C A C C T G C T A C C A T T A T A T A T T A T T A C C T A C A G A C A C A T T A T T A T G A C T C A A A A C A T C A A A A C A T A C C T T G C G T G C A G C A C C 3480 (NAS)

Y I P R M S D T M I N Y R T T N S T P S I D D R S S H S E S T T P I I S I E T N 680 (AAS)
 A T A T T A C C A A A T A T C A A A T A T C A A A T A C A A T A C C C C C T A A T A G A T A C C C T T C C T C A A C A T G A T C A T C C C A T C C A T C T C C A T A C A A A A 3600 (NAS)

A U C D N T P S I D T D P P E Y R S S D H A T P N I N P D K S S K N U T A D S I 720 (AAS)
 G C G T A T C A T A A T A C A C C T C C A T T A C A T A C G A T C C C A G A T A T C A T C T T C T A C C A T C G A C T C C T A T A T A T A T G C C T A C A A A T C C A A A A T T G T A D G C T G A T T A T 3720 (NAS)

L D D L P L P D I T H Q S P T D T S D V S K D I P H I H S R Q T N S S L G C M D 760 (AAS)
 T C T A C A C C T C C A C T T C C T A C T A A C C A T C A A T C C T A C C A C A C T C T G A T G T T C A A A G A T A T T C C A C A C A T A C A C T C T C T C A C A T A T T C C A C T T G C T G T A T G A 3840 (NAS)

P S M V L T T T T K S K V R S L E D N E T E I E V S R D T M N N K N M R S L E P P 800 (AAS)
 T G T T A T T A T T C T A C T A C C A A A A T C A C A A A G A T C A T T A G A C A T A T G A A A T T C A G C T A T C C C A G A C A C A T G A A T A T A C A A T A T A C A A T C C A C C A C C 3960 (NAS)

S K K R I M L I A A I X G U K S I M P U R T T L R Y D E A I T Y N K D N F E K 840 (AAS)
 A G A T C A A A A C C A T A A T T A T T A T T C A C A A T A A A G A C T C A A T C A A A C A C T C A A C A C T T A A G A T A T C A T C A A G A T A T T A T A T A A A C A C A A A A A A A 4080 (NAS)

P R I V E A Y H K E I S O I I K N N T M D T N K Y Y D R N D I D P K X V I N S H 880 (AAS)
 A G C A G A T G T T A A C C T A T C A T A A A A A T A A C C A C T A T T C A A A A T A C A C T T G C G A T C A A A A A T A T T A T A T A C A A T G A C A T C C A A A A A T A A A C T A A T 4200 (NAS)

F I F N K K P D G T H K A R F V A R G D I Q H P D T Y D S D M Q S N T V H H Y A 920 (AAS)
 G T T A T A T T A A C A A A C C T A G A C A C A A A A G C A T T G C T T C C A G A C C G A C A T C A A C A C C C G A T A T A T G A T T C A T A T C A A T C A A C C T A C C T A C A C T A T C C 4320 (NAS)

L M T S L S I A L D N D Y Y I T O L D I S S A Y L Y A D I K E E L Y I R P P P H 960 (AAS)
 A C T A T A C C T C A C T C A T C C C A T T A C A C A C A C A T A T A T A T A C A C A C T C A C A T A T C C T C C T A C T A T A T C C T A T A C C A T A A A C A A C A T A T A C A T A C A C C T C C A C A C A 4440 (NAS)

L G L N D F L L R L R K S L Y G L K Q S C A M W Y E T I K S Y L I M C C D M O E 1000 (AAS)
 T T T A C T T C A T A T A A T A C T A C C T T T C A A A A T C A C T C A T G C T T C A A A A G T G C C A A A T G C T A T G A A C C A T A A A T A T A T T A T A A T T A T T T C C A C A T C A C A 4560 (NAS)

V R C M S C V F P R N S Q V T I C L F V D D M I L F S K D L N A M K K I I T T L K 1040 (AAS)
 A C T C C G A T G C T A C C T A T T A G A C A T G C A A T T A C A A T T T C T T A T C C T G C A T G A T A T A T T A T T A T C A C A A C A C T A T T A T T A C A A A A C A A A A A A A A A A A A A 4680 (NAS)

K O Y D T K I I N L G E S D N E I Q Y D I L G L E I X Y O R S K Y M K L G M E A 1080 (AAS)
 C A A C A A T A C G A T A C A A A C A T A A A T C C G C T G A A A C T A T A C C A A A T T C A C T A C C A C A T A C T T C A T T A G A C A T A A A T C A A A G C A A C A T A C A A A T A G C T A T G C A A A 4800 (NAS)

S L T E K I P K I N V P L N P K K X L R A P G Q P G H Y I D D E L E I D E D 1120 (AAS)
 A C C T T C A C A C A A A A C T A C C A A C T A C C T T T C A A C C A A A A G A A A A C T A G A C C T C A G C T C A A C C A C C A T T A T A T A T A C A C A C A C A T G C C A A C A T A C A T G A C A 4920 (NAS)

F Y R E A V H E M G K L I C L A S Y V G Y K F R F D L L Y Y I N T L L N H I L F 1160 (AAS)
 T G A T A C A A G A C A A C T A C T A A A T C A A A A T T A C T T A C C T A C T A T A T T G A T A T A A T T A G A T T G A C T A C T A T A C A C A C A C A T G C C A A C A T A C T A T T 5040 (NAS)

P S R Q U L I N F Y E L I O F H W D T R D K G L I W H K N A P T R P D N A L V A 1200 (AAS)
 C C C T T A G C A C T T T A G A C A T A C A T A C A T T A T A C A T T C A T T G C G A C A C A G A T A A C A T A T A T G C A C A A A A C A A A C T A C A C A C C A T A T A A A C T A G C C 5160 (NAS)

I S D A S Y G N O P Y Y K S O I G N I F L L N G X V I G C K S T K A S L T C T S 1240 (AAS)
 A T A A C C A C C T C A T A T A T A C C A C A T A T T G T A C A T T T T C C A C T A C A C C A A A A C T A G A C C T A T T T C T A C T A A C C A A A A C T A T T G C A A A G C C T G T T A T A C A T C A C T T C 5280 (NAS)

T T E A E I H A U S E A I P L L N N L S H L V Q E L N K K P I I N G L L T D S R 1280 (AAS)
 A A C T A C A G A C A A A T A C A C C G C T A C T G A C T A T T C C C T A T T C A T A A C C T C A C T A C C T T C T G C A G A C T T A C A C A A A C A A T T A T A A A G C T A C T A C T A C T A C T A G 5400 (NAS)

S T I S I I K S T N E E K F K N R F F C T K A M R L R D E U S C N N L Y V Y Y I 1320 (AAS)
 A T C A C A T C A T A T A T A T A C T T A C A A T A C A G A A A T T A G A A C A G A T T T T T G C A C A A G C A A T G A C A C T A G A C A T G A T A C A T A T A T T A T A C T A T A C A T 5520 (NAS)

E T K N I A D U T N K P L P I N T F R L L T N X W I N * 1349 (AAS)
 C G A C C A A G A A C A T T C T G A T G T A T C A A A A C C T T C C G A T A A A A A C T T T A A C T A T T A C T A C A A A T G A T T A T T A T T A C T A T T A C T A T T A T T G C T G T A T C T T G G A T A 5640 (NAS)

A A A T C A C A T A C T A C T A C T A T T A C T T A C T A T A T A T A C A T A C C G T C T A G A A T A C C A A A T G A T A C A A A T C A T C A T A A A T A G C A A C T A A A G C A A 5760

G C A T T A T A T C A T A G C A T C A T A T T A C A T A A A A T G A T A T A T A T T A T A G A A T T G T A G A A T C A A T C C C T T T A T G C A T T C T A A A T C C T G A C A G A C T 5880

T T A C T A T A C T A C A C C A T A A T A T T G C C T T A T A A A A T G C A T C C A A A T T A C A T A A A A T C C A A T T C T T C A 5962

Figure 2. Complete nucleotide sequence of the class II Ty element, Ty1=17 from chromosome III of *S. cerevisiae*. Regions of interest marked are discussed in the text.

distal side, Ty 1-17 is preceded by a tRNA₃^{glu} gene and two solo delta sequences (J.R. Warmington et al., in preparation). Transfer RNA genes are frequently associated with Ty and other repetitive elements in yeast (23).

Figure 3. DIAGON dot matrix (24) comparisons of the nucleotide sequences of Ty elements. Dots represent the perfect matching of 21 nucleotides with a probability of chance occurrence of 10^{-4} . The numbers indicate DNA sequence position in kb. **a.** Comparison of two class I Ty elements; Ty 912 (abscissa) and Ty 109 (ordinate). **b.** Comparison of the class II element, Ty 1-17 (abscissa) with the class I element, Ty 912 (ordinate). The "echoes" in the bottom left and top right-hand corners of the diagram are caused by the directly repeated, terminal delta sequences.

Sequence organization of Ty 1-17

The complete nucleotide sequence of Ty 1-17 is presented in Fig. 2. The element consists of an internal (epsilon) sequence of 5297 bp flanked by two identical, directly repeated, delta sequences of 332 bp each. The element is bracketed by a 5 bp direct repeat 5' AAATC 3' which represents the duplication of the chromosomal target sequence produced by the transposition event which inserted Ty 1-17 into chromosome III.

The nucleotide sequences of class I and class II Ty elements were compared using DIAGON dot matrix analysis (24). This revealed complete sequence homology between the epsilon regions of the two class I elements, Ty 109 (ref. 15) and Ty 912 (ref. 14), but some divergence in their delta sequences (Fig. 3a). A comparison of the sequence of the class II element Ty 1-17 with that of Ty 912 is shown in Fig. 3b. In contrast to the previous comparison, these two elements show complete homology in their delta sequences but considerable divergence in their epsilon sequences. The regions of heterogeneity correspond to those previously detected by analysis of heteroduplexes formed between DNA molecules of class I and class II elements (12 and Fig. 1). We next examined whether this heterogeneity is reflected in the amino acid sequence of the ORFs which the two classes of element contained.

Table 1. A comparison of the *TYA* and *TYB* ORFs in class II and class I elements.

		Start	Nucl- eotide position	Stop	Nucl- eotide position	Length amino acid	mwt (kd)
Class II	<i>TYA</i> 17	ATG	292	TAA	1608	438	49.8
Class I	<i>TYA</i> 109	ATG	291	TGA	1613	440	49.1
	<i>TYA</i> 912	ATG	294	TGA	1616	440	48.9
	<i>TYA</i> 15	ATG	299	TGA	1619	440	49.1
Class II	<i>TYB</i> 17	GCG	1562	TAG	5608	1349	154.1
Class I	<i>TYB</i> 109	ACA	1573	TAG	5559	1328	151.1
	<i>TYB</i> 912	ACA	1576	TAG	5562	1329	151.2

Open reading-frames

Fig. 2 shows *Ty* 1-17 to contain two large open reading-frames, *TYA*17 and *TYB*17 (13) which overlap one another and are out of phase by a +1 frameshift. *TYA*17 is initiated with an ATG codon at position 292 and extends for 438 amino acids, terminating with TAA termination codon at position 1608. The termination codon for the *TYA* open reading-frame of class I elements is TGA. The putative protein products from both class I and II elements are similar in both size and amino acid composition (Table 1). A more detailed comparison of the *TYA* open reading-frames in the two classes of element has been published previously (13).

The second open reading-frame, *TYB*17 begins with a GCG (alanine) codon at position 1562, although there is an in-frame ATG (methionine) immediately adjacent to this. *TYB*17 terminates at a TAG codon at position 5608. The class I elements also terminate with a TAG but encode a potential protein product some 20 amino acids shorter than the *TYB*17 product (Table 1).

The *TYB*17 amino acid sequence was compared with that of *TYB*912 using the DIAGON dot matrix sequence comparison programme (24). This comparison (Fig. 4b) demonstrated that the two proteins share two regions of amino acid sequence homology separated by an heterogeneous segment. In contrast, a comparison of the amino acid sequences of *TYB*109 and *TYB*912 (Fig. 4a) show near-perfect homology between the two proteins. It may be that the two regions of homogeneity, which we call *TYB*' and *TYB*", represent conserved functional domains within this protein. *TyB*' extends from amino acid 1 to ca. 520 and *TyB*" from ca. amino acid 740 to the end of the *TyB* ORF. The amino acid compositions of *TYB*' and *TYB*" from class I and class II elements are compared in Table 2.

Figure 4. DIAGON dot matrix comparisons of the amino acid sequences of TyB ORFs. Dots represent proportional matching over a span length of 29 amino acids; the probability of homologies occurring by chance is 10^{-4} . The numbers indicate amino acid position within the TYB ORF. **a.** Comparison between two class I elements, Ty 912 (abscissa) and Ty 109 (ordinate). **b.** Comparison between the class II element, Ty1-17 (abscissa) and Ty 912 (ordinate).

Specific sequence motifs conserved in both classes of element

The transcription of yeast Ty elements begins within the 5' delta 90-100 bp upstream from the delta/epsilon junction and terminates at a site ca. 300 bp into the 3'

Table 2. The amino acid composition of the two conserved domains of the TYB ORF in the two classes of Ty element.

	Non-polar							Polar							Acid		Basic		MW		
	A	V	L	I	P	F	W	M	G	S	T	C	Y	N	Q	D	E	K	R	H	(kd)
a)																					
TYB'17	26	21	54	36	25	20	4	3	19	45	39	7	23	37	25	31	24	32	23	27	59.5
TYB'109	29	25	52	41	26	23	6	1	19	52	40	5	23	34	19	34	18	27	22	25	58.8
TYB'912	29	24	49	40	25	23	4	1	20	50	42	6	24	36	20	35	20	22	25	25	59.0
b)																					
TYB''17	22	23	62	51	22	15	7	15	25	42	39	5	33	42	20	42	33	67	27	15	70.4
TYB''109	24	23	59	53	24	15	7	15	26	42	41	4	34	37	21	37	34	66	28	17	70.5
TYB''912	23	23	60	53	23	15	7	16	27	42	36	4	34	39	22	35	40	66	29	13	70.4

a) TYB' = amino acids 1 to 520. b) TYB'' = amino acids 741 to 1349 (Ty1-17); 721 to 1328 (Ty109); 721 to 1329 (Ty912).

Table 3. Conserved motifs within the nucleotide sequences of both classes of Ty element.

a.									
		1		2		3		4	
	Ty1-17	167	TATAAAA	241	GAGGA	5882	TAGT	5930	CAACAA
	Ty109	162	TATAAAA	238	GAGGA	5831	TAGT	5880	CAACAA
	Ty912	167	TATAAAA	241	GAGGA	5837	TAGT	5873	CAACAA
b.									
		1		2		3			
		...	...				...		
	Ty1-17	648	GCGTATTAT-CAACC-GGAC	789	ACCACCACACACTTT	815	AAA-ACTTTTCTACA-TGGG		
	Ty109	648	ATGTACTTT-CCACC-TGGG	800	ACCACCAC-CAATGT	826	ATG-ACTTT-CCAAATTGGG		
	Ty912	651	ATGTACTTT-CCACC-TGGG	803	ACCACCAC-CAATGT	828	ATG-ACTTT-CCAAATTGGG		
	Ty1-15	649	ATGTACTTT-CCACC-TGGG	800	ACCACCAC-CAATGT	826	ATG-ACTTT-CCAAATTGGG		
c.									
	Ty1-17	393	CGTTTCAGCTTCCA	5744	TGGAAGCTGAAACG				
	Ty109	392	CGTTTCAGCTTCCA	5689	AGGAAGCTGAAGTG				
	Ty912	395	CGTTTCAGCTTCCA	5698	TGGAAGCTGAAACG				
d.									
	TY1-17	333	TGGTAGCGCCTATGCTTCGGTTAC						
	Ty109	332	TGGTAGCGCCTGTGCTTCGGTTAC						
	Ty912	335	TGGTAGCGCCTGTGCTTCGGTTAC						
	Ty1-15	340	TGGTAGCGCCTGTGCTTCGGTTAC						
	tRNA _i ^{met}	3	ACCAUCGCGG---CGA-GCCAAAG						

a. 1, TATA box; 2, transcription initiation site; 3, beginning of transcriptional termination signal. 4, 3' end of R N A transcript. **b.** 1, upstream diploid regulation site; 2, potential enhancer sequence; 3, downstream diploid regulation site. **c.** reverse complement sequences **d.** complementarity at the 5' delta/epsilon junction with the 3' end of yeast tRNA_i^{met}. Numbers indicate nucleotide position within the sequence.

delta sequence (25). Both classes of Ty element contain conserved promoter and termination sequences at equivalent positions within the delta elements (see Table 3). Since the transcriptional initiation site precedes the termination site, the latter must be ignored by the RNA polymerase complex in the 5' delta, but recognized by it in the 3'

delta sequence. The large transcript produced represents the template for the action of reverse transcriptase during transposition (10); however, it is missing the extreme 5' and 3' ends of the element. This problem is dealt with at greater length below.

The process of DNA synthesis from an RNA template requires a primer for its initiation. In retroviral replication this requirement is fulfilled by the annealing of a tRNA molecule to a sequence within the positive strand of the virus complementary to the 3' end of the transfer RNA (26). A similar mechanism is involved in the reverse transcription of cauliflower mosaic virus (5). Both class I and class II yeast Ty elements contain a sequence complementary to that of yeast tRNA_i^{met} (27) at the 5' delta/epsilon boundary. This is a highly conserved region where the three published sequences of class I elements are identical and that of Ty 1-17 differs from them by only a single nucleotide, which is probably not involved in base-pairing with the tRNA_i^{met} (Table 3d). In retroviruses (4,28) and cauliflower mosaic virus (29), synthesis of the second (positive) DNA strand is primed by a G-rich segment of the positive RNA strand which is resistant to the action of RNase H. Ty 1-17, Ty912 and Ty109 all contain the sequence 5'TGGGTGGT3' at the extreme 3' end of their epsilon regions, immediately in front of the junction with the 3' delta element.

The transcripts of both class I and class II elements contain a pair of sequences, 14 nucleotides long, which are the reverse complement of one another. One sequence occurs at the 5' end of the epsilon region 152-156 nucleotides from the start of the transcript, its reverse complement is contained within the transcribed portion of the delta sequence, 185-190 nucleotides from the 3' end of the RNA molecule (Table 3c).

The transcription of yeast Ty elements is under mating-type control (25). Insertion of either class of Ty element into a yeast chromosome upstream of a structural gene may place the expression of that gene under mating-type control as well; an effect known as a ROAM mutation (30,31,32,33,34). Both class I and class II Ty elements contain sequences homologous to the so-called "diploid control site" found in the MAT locus. This is a sequence which lies between the two transcribed portions of the MAT α gene and regulates its expression (35). The class I elements contain two sequences, each showing a 16/18 bp match with the MAT α diploid regulation site. In the class II element, Ty 1-17, the equivalent upstream sequence shows only a 9/18 bp homology with the diploid regulation site, whereas the downstream sequence gives a 16/18 bp match. It is noteworthy that the sequence which is conserved in both classes of element is situated immediately downstream of an enhancer-like element (36). These relationships are described in Table 3b.

DISCUSSION

The sequence analysis of the class II yeast transposon, Ty 1-17, has demonstrated that it, like the class I elements, has a retrotransposon type of genome arrangement.

Table 4. Homology within the amino acid sequences of the two O R F s in regions showing similarity to functional domains within bacterial transposons and retroviruses.

a.

TYA17 236 ILEINYADILTVLCKSVSKM 300 VSDRLACQLILKGLSGDFKYLRNQYRTKTNMKLSQL
TYA109 240 ILSVDYTDIMKILSKSIEKM 305 INNKVACQLIMRGLSGEYKFLRYTRHRHLNMTVAEL
TYA912 240 ILSVDYTDIMKILSKSIEKM 305 INNKVACQLIMRGLSGEYKFLRYTRHRHLNMTVAEL
TYA15 241 ILSVDYTDIMKILSKSIEKM 304 INNKVACQLIMRGLSGEYKFLRYTRHRHLNMTVAEL

b.

TYB17 28 PDHLLIDSGRSQTLVRS
TYB109 28 PGHLLLD SGASRTLIRS
TYB912 28 PGHLLLD SGASRTLIRS
TYB15 28 PGHLLLD SGASRTLIRS

c.

M3 M4

TYB17 305 VIQMDRGSEYTNKTLHKFFTNRG 341 HGVAERLNRTLLND
TYB109 305 VIQWTVVLSILSRTLHKFLEKNG 341 HGVAERLNRTLLDE
TYB912 305 VIQMDRGSEYTNRTLHKFLEKNG 341 HGVAERLNRTLLDD

R2

876 VINSMFIFNKKRDGTHKARFVA
857 VINSMFIFNKKRDGTHKARFVA
856 VINSMFIFNKKRDGTHKRRFVA

I II III

946 YADIKEELYIRP 990 SYLINCCDMQEVRGWSCVFKNSQVTICLFVDDMILFSKD
927 YADIKEELYIRP 970 IIPDKTVWYGRSRGWSCVFKNSQVTICLFVDDMILFSKD
926 YADIKEELYIRP 970 SYLIQCGMEEVRGWSCVFKNSQVTICLFVDDMVLFSKN

a. DNA binding - site homology within TYA (see ref. 14, 15). **b.** Region showing homology to retroviral protease (42). **c.** Region showing homology to retroviral reverse transcriptase (14, 41). Numbers indicate amino acid position within the sequence of the relevant O R F. Amino acids which show neither homology nor conservation of chemical characteristics are underlined.

Although there is considerable sequence divergence between the two classes of element, the amino acid sequence of both open reading-frames is highly conserved. The first open reading-frame, TYA, encodes a protein of ca. 50,000 MW which has been demonstrated to

have nucleic acid binding activity (37). The second, and much larger, open reading-frame (TYB) has two regions of amino acid sequence (TYB', TYB'') which are highly conserved between the two classes of elements. It is possible that the two conserved regions represent two functional domains within the TYB protein. By analogy with the organization of the retroviral genome TYA, TYB' and TYB'' would be equivalent to the gag gene, and the protease and reverse transcriptase domains of the pol gene, respectively (4,38). There are a number of features of the amino acid sequence of these regions which support this contention.

The TYA protein is a basic, proline-rich polypeptide (13) which, like the gag proteins of retroviruses, has nucleic acid binding activity (37). Hauber et al. (15) have pointed out a region of homology between the TYA₁₀₉ and the resolvase protein tnpA of the bacterial transposon Tn3 (38), while Clare and Farabaugh (14) have noted an adjacent sequence similar to the consensus for bacterial DNA-binding proteins (40). Both of these sequences are highly conserved between the two classes of yeast transposon (Table 4a). In collaboration with Dr. J.R. Wooton of the University of Leeds, we have found considerable structural homologies between TYA₁₇ and both gag and resolvase proteins (to be published elsewhere).

The TyB' region of the second ORF contains a region which shows homology to the amino acid sequence of the retrovirus protease (see Table 4b). This sequence of 17 amino acids is identical in all three class I elements sequenced so far. The class II element (Ty₁₋₁₇) shows four amino acid changes in this sequence, two of which are chemically conservative alterations.

The TyB'' domain is basic in character and shows a number of regions of similarity to the retroviral and cauliflower mosaic virus reverse transcriptases (14,39,42). All of these regions are highly conserved between the class I and class II elements, there being no more sequence divergence between classes than there is between the two class I elements (Table 4c). Two of the regions described by Clare and Farabaugh (14) as resembling reverse transcriptase (M3 and M4; see ref. 40) lie within the TyB' region, but their homology with the amino acid sequence of the enzyme is poor.

The interstitial region between TyB' and TyB'' is polar in character. Both class I and class II Ty elements contain the motif SKKRS/I, repeated twice, near the boundary of the interstitial region and TyB''. These amino acid sequences represent potential sites for the proteolytic cleavage of the TyB polypeptide.

Boeke et al. (10) have provided a convincing demonstration that Ty elements are transposed via an RNA intermediate. The sequence analysis presented here suggests regions of the transcript which might be involved in priming the synthesis of the first and second DNA strands. The problem remains as to how a complete DNA copy of a Ty element can be generated by reverse transcription when the RNA transcript is initiated

Figure 5. Model for the generation of a DNA copy of the Ty transcript. The first DNA strand (→) is initiated from a tRNA_i^{met} primer. The U5 region of the 5' delta is copied initially since this is absent from the 3' delta sequence of the transcript. Within the R region, there is a switch in the template strand used from the 5' to the 3' delta. DNA synthesis then continues around the loop but a second switch in template must be made in order to provide a DNA copy complete with a 5' U3 delta region. These switches in template strand utilization are facilitated by base-pairing between the two reverse complement sequences. The synthesis of the second DNA strand (←→) is initiated from a purine-rich (and thus nuclease-resistant) oligoribonucleotide at the junction of the epsilon and right-hand delta sequences.

and terminated within the delta sequences. Boeke et al. (10) showed that, in the transposed copy of TyH3, the 5' portion of the left-hand delta (designated U3) is derived from the delta sequence of the RNA transcript, while the 3' portion of the right-hand delta (designated U5) is derived from the 5' delta sequence of the transcript. We propose that the reverse complement sequences which are found in both class I and class II elements could be used to bring the 5' and 3' ends of the transcript into suitable juxtaposition for first and second strand DNA synthesis to replace the missing portions of the element in the manner described by Boeke et al. (10). This model is outlined in Fig. 5. In retroviruses, a similar role may be played by small inverted repeat sequences found wholly within the LTR (4).

The sequence analysis of Ty 1-17 has enabled us to identify conserved sequences which are found in both classes of Ty element and which are likely to have functional significance. We have suggested a number of ways in which these sequences could be involved in the mechanism of Ty transposition or in the way these transposons are able to

exert genetic effects within yeast. All of these suggestions are open to experimental test.

ACKNOWLEDGEMENTS

This work was supported by SERC grant no. GR/C34656 (to SGO and R.W. Davies), by NIH grant no. GM21510 (to CSN) and by a NATO collaborative Award (to SGO and CSN). Wayne Davies is thanked for his interest and encouragement. We are also grateful to Phil Farabaugh, Horst Feldmann and Alan Kingsman for the communication of data prior to publication, and to John Wooton for helpful discussions.

*To whom correspondence should be addressed

REFERENCES

1. McClintock, B. (1952) Cold Spring Harbor Symp. Quant. Biol. 16,13-42.
2. McClintock, B. (1957) Cold Spring Harbor Symp. Quant. Biol. 21,197-206.
3. Shapiro, J.A., Ed. (1983) Mobile Genetic Elements, Academic Press, New York.
4. Varmus, H.E. (1983) In: Mobile Genetic Elements, Shapiro, J.A., ed., Academic Press, New York pp. 411-503.
5. Hohn, T., Hohn, B. and Pfeiffer, P. (1985) Trends in Biochem. Sci. 205-209.
6. Rubin, G.M. (1983) In: Mobile Genetic Elements, Shapiro, J.A., ed., Academic Press, New York pp. 329-361.
7. Williamson, V.M. (1983) Int. Rev. Cytol. 83,1-25.
8. Heffron, F. (1983) In: Mobile Genetic Elements, Shapiro, J.A., ed., Academic Press, New York pp. 223-260.
9. Baltimore, D. (1985) Cell 40, 481-482.
10. Boeke, J.D., Garfinkel, D.J., Styles, C.A. and Fink, G.R. (1985) Cell 40,491-500.
11. Cameron, J.R., Loh, E.Y. and Davis, R.W. (1979) Cell 16,739-751.
12. Kingsman, A.J., Gimlich, R.L., Clarke, L., Chinault, A.C. and Carbon, J. (1981) J. Molec. Biol. 145,619-632.
13. Fulton, A.M., Mellor, J., Dobson, M.J., Chester, J., Warmington, J.R., Indge, K.J., Oliver, S.G., de la Paz, P., Wilson, W., Kingsman, A.J. and Kingsman, S.M. (1985) Nuc. Acids Res. 13,4097-4112.
14. Clare, J.J. and Farabaugh, P.H. (1985) Proc. Natl. Acad. Sci. U.S.A. 82,2829-2833.
15. Hauber, J., Nelback-Hochstetter, P. and Feldmann, H. (1985) Nucleic Acids Res. 13,2745-2758.
16. Deininger, P.L. (1983) Anal. Biochem. 129,216-223.
17. Newlon, C.S., Devenish, R.J., Lipchitz, L.R. (1982) Recent Adv. in Yeast Molec. Biol. 1,52-58.
18. Messing, J. (1983) Methods in Enzymol. 101,20-78.
19. Biggin, M.D., Gibson, T.J. and Hong, G.F. (1983) Proc. Natl. Acad. Sci. U.S.A. 80,3963-3965.
20. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. U.S.A. 74,5463-5467.
21. Strathern, J.N., Newlon, C.S., Herskowitz, I. and Hicks, J.B. (1979) Cell 18,309-319.
22. Andreadis, A., Hsu, Yun-Pung, Kohlhaw, G.B. and Schimmel, P. (1982) Cell 31,319-325.
23. Eigel, A., and Feldmann, H. (1982) EMBO J. 10,1245-1250.
24. Staden, R. and McLachlan, L. (1982) N.A.R. 10,141- .

25. Elder, R.T., Loh, E.Y., and Davis, R.W. (1983) *Proc. Natl. Acad. Sci. U.S.A.* 80,2432-2436.
26. Taylor, J.M. (1977) *Biochim. Biophys. Acta.* 473,57-71.
27. Venegas, A., Gonzalez, E., Bull, P., and Valenzuela, P. (1982) *Nucleic Acids Res.* 10,1093-1096.
28. Mitra, S.W., Chow, M., Champoux, J. and Baltimore, D. (1982) *J. Biol. Chem.* 257,5983-5986.
29. Smith, J.K., Cywinski, A., and Taylor, J.M. (1984) *J. Virol.* 52,314-319.
30. Dubois, E., Jacobs, E., and Jauniaux, J.C. (1982) *EMBO J.* 1,1133-1139.
31. Errede, B., Cardillo, T.S., Wever, G., and Sherman, F. (1980) *Cold Spring Harbor Symp. Quant. Biol* 45,593-602.
32. Kosiba, B.E., Errede, B., Cardillo, T.S. and Sherman, F. (1982) *Rec. Adv. Yeast Mol. Biol.* 1,156-172.
33. Williamson, V.M., Cox, D., Young, E.T., Russell, D.W. and Smith, M. (1983) *Molec. Cell. Biol.* 3,20-31.
34. Taguchi, A.K.W., Ciriacy, M. and Young, E.T. (1984) *Molec. Cell. Biol.* 4,61-68.
35. Siliciano, P.G. and Tatchell, K. (1984) *Cell* 37,969-978.
36. Khoury, G. and Gruss, P. (1983) *Cell* 33,313-314.
37. Kingsman, A.J., personal communication.
38. Shinnick, T.M., Lerner, R.A. and Sutcliffe, J.G. (1981) *Nature* 293,543-548.
39. Heffron, F., McCarthy, B.J., Ohtsubo, H. and Ohtsubo, E. (1979) *Cell* 18,1153-1163.
40. Pabo, C.O. and Sauer, R.T. (1984) *Ann. Rev. Biochem.* 53,293-321.
41. Patarca, R., Haseltine, W.A. (1984) *Nature* 309,728.
42. Toh, H., Kikuno, R., Hayashida, H., Miyata, T., Kugiyama, W., Inouye, S., Yuki, S. and Saigo, K. (1985) *EMBO J.* 4,1267-1272.