Volume 13 Number 11 1985 Nucleic Acids Research

Variants within the yeast Ty sequence family encode a class of structurally conserved proteins

Alexandra M.Fulton, Jane Mellor, Melanie J.Dobson, John Chester, John R.Warmington[1], Keith J.Indge[1], Stephen G.Oliver[1], Patricia de la Paz[2], Wilma Wilson, Alan J.Kingsman* and Susan M.Kingsman

Departments of Biochemistry and [2]Molecular Biophysics, University of Oxford, South Parks Road, Oxford OX1 3QU, and [1]Department of Biochemistry and Applied Molecular Biology, UMIST, P.O. Box 88, Manchester M60 1QD, UK

Received 19 April 1985; Revised and Accepted 10 May 1985

ABSTRACT

The Ty transposable elements of Saccharomyces cerevisiae form a heterogeneous family within which two broad structural classes (I and II) exist. The two classes differ by two large substitutions and many restriction sites. We show that, like class I elements a class II element, Ty1-17, also appears to contain at least two major protein coding regions, designated TYA and TYB, and the organisational relationship of these regions has been conserved. The TYA genes of both classes encode proteins, designated p1 proteins, with an approximate molecular weight of 50 Kd and, despite considerable variation between the TYA regions at the DNA level, the structures of these proteins are remarkably similar. These observations strongly suggest that the p1 proteins of Ty elements are functionally significant and that they have been subject to selection.

INTRODUCTION

The yeast Ty element is a member of a class of eukaryotic transposons that resembles retroviral proviruses in both structure and functional expression strategies (1,2,3). It is unusual, however, in that the sequence family that comprises some 30-35 copies of the element in most laboratory strains exhibits marked heterogeneity (4). Ty elements isolated to date appear to fall into two major classes (I and II) that differ by two large substitution loops in EM heteroduplex analysis (4, Figure 1) . Within these classes there are minor restriction site variations. The superficially similar copia-like elements in Drosophila do not show such variation (5). Different Ty elements may encode different, perhaps complementing activities, some elements may be functional while others are not or class I and II elements may be diverging into independent elements. In order to address these alternatives we are analysing the structure and function of different elements; for example if some Ty elements are non-functional we might expect loss of an ordered genetic organisation in those elements.

Ty elements are about 5.9 kb in length with direct 340 bp terminal repeats, called delta sequences, flanking an internal 5.2 kb epsilon region (6). A major 5.7 kb transcript starts about 240 bp into the left or 5' delta and ends 290 bp in the 3' delta (7; Figure la). A typical class I element, Tyl-15 (4; Figure la) contains two open reading frames, previously designated ORF1 and ORF2 (2). These have been shown to encode proteins *in vivo* (2,8) and have therefore been given the gene designations *TYA*15 and *TYB*15. *TYA*15 extends from nucleotide 299 to 1619 and encodes a 50 Kd basic protein, pl(Tyl-15), *in vivo* (8). *TYB*15 begins with an ACA (threonine) codon at nucleotide 1581 and extends for at least 500 bp in Tyl-15. The entire nucleotide sequence of an element very similar to Tyl-15 has recently been determined by Hauber *et al* (1) and in this element TYB extends to the end of the epsilon region. *TYA*15 and *TYB*15 overlap by 38 nucleotides and are in different reading frames. *TYB*15 is expressed as a fusion protein, p3(Tyl-15) (Mellor *et al*., manuscript in preparation) with most, if not all, of *TYA*15 at the N-terminus (2). This fusion involves a frameshift in the region of the overlap of the ORFs (2). The region of Tyl-15 which contains *TYA* and the overlap junction between *TYA* and *TYB* shows no homology in a heteroduplex with a typical class II element, Tyl-17 (4; Figure la). This might imply either that *TYA* is not functionally significant to Ty elements and is therefore not conserved, or that the different Ty elements encode entirely different proteins which are important but have no functional correlation. Another possibility is that not all Ty elements have conserved open reading frames, for example the class II elements could be non-functional. Alternatively, the lack of homology at the DNA level may not be reflected in primary or secondary structural features of the proteins.

In this paper we present an analysis of the gene expression of this region of non-homology in a class II element, Tyl-17. The aim is to show the relationship between class I and class II Ty elements and to infer biological significance of the *TYA* regions from this comparison. We show that, as in Tyl-15 (8) the TYA region of Tyl-17 encodes a protein *in vivo*. We also demonstrate that, despite considerable variation between class I and class II elements at the DNA sequence level, the encoded proteins from the *TYA* regions show extensive structural homology and the relationship of the *TYA* and *TYB* regions has been

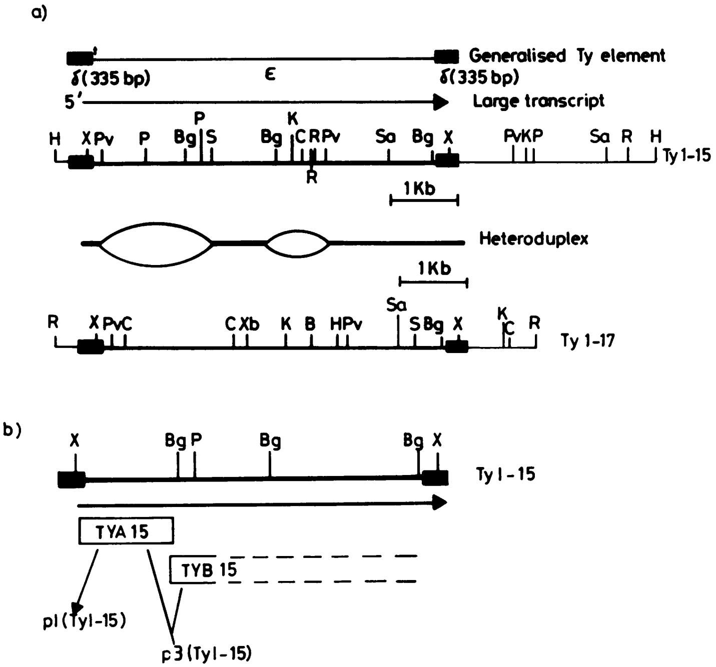

Figure 1. a. Restriction maps and gross sequence relationships of Tyl-15 and Tyl-17. The diagram at the top shows the general organisation of a Ty element and its major transcript. A 9.6 Kb HindIII fragment containing Tyl-15 and a 7.2 Kb EcoRI fragment containing Tyl-17 are shown with a schematic diagram of the heteroduplex formed between the two elements (4). b. The genetic organisation of Tyl-15. The long arrow marks the major transcript and the two major open reading frames *TYA*15 and *TYB*15 are shown. The two primary translation products of Tyl-15, pl(Tyl-15) and p3(Tyl-15) are also marked. H = HindIII; X = XhoI; Pv = PvuII; P = PstI; Bg = BglII; S = SalI; K = KpnI; C = ClaI; R = EcoRI; Sa = SacI; Xb = XbaI; B = BamHI. Closed boxes = the long terminal repeat delta sequences.

conserved. The data strongly suggest that both classes of Ty elements encode functional proteins that have been subject to selection.

MATERIALS AND METHODS

Bacterial and yeast strains and media

E.coli strain AKEC28 (C600 *thr*C *leu*B6 *thy*A *trp*C1117 *hsd*Rk *hsd*Mk) was used for plasmid manipulation and preparation. *Saccharomyces cerevisiae* strain MD40-4c = *ura*2 *trp*1 *leu*2-3 *leu*2-

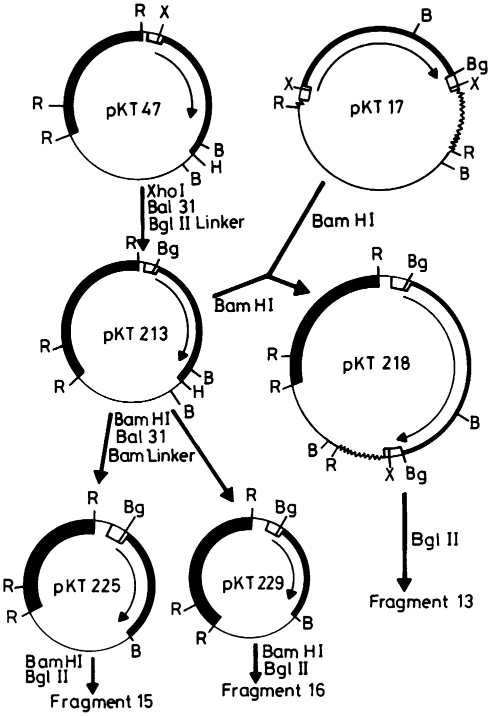

Figure 2. The construction of Ty1-17 fragments 13, 15 and 16. Closed region = 2u origin of replication, REP3 and the LEU2 gene; open box = delta sequence; thin line = pBR322 sequences; wavy line = yeast flanking sequences. Arrow marks the direction of Ty transcription. X = XhoI; R = EcoRI; B = BamHI; Bg = BglII; H = HindIII.

112 *his*3-11 *his*3-15. *E.coli* media were prepared according to Miller (9). Yeast media were prepared according to Hawthorne and Mortimer (10).

DNA isolation

Plasmid DNA was isolated as described by Chinault and Carbon (11). Mini preparations of plasmid DNA were according to Holmes and Quigley (12).

Yeast transformation

The method of Hinnen *et al.*(13) was used.

Enzymes and fragment purification

Enzymes were purchased from BRL and used according to the manufacturers instructions. DNA fragments were purified from agarose gels by the method of Tabak and Flavell (14).

Construction of subfragments 13, 15 and 16 from Tyl-17

Plasmid pKT47 was the starting molecule for these constructions (Figure 2). It is a 2u-based plasmid containing the EcoRI:HindIII fragment corresponding to the left 'half' of Tyl-17 (Figure 1a). It was cleaved with XhoI and subjected to a very short Bal31 treatment. BglII linkers, CAAAAGATCTTTTG, were then added to produce pKT213. To produce a plasmid containing fragment 13 pKT213 was cleaved with BamHI and ligated in the presence of a BamHI digest of pKT17. Plasmid pKT17 is a pAT153 derivative containing the 7.2 kb EcoRI fragment containing Tyl-17 (Figure 1a). A recombinant, designated pKT218, containing a reconstructed Tyl-17 element but with the BglII site at the 5' end was obtained. A BglII digest of pKT218 was the source of fragment 13. Plasmid pKT213 was also cleaved with BamHI and subjected to Bal31 exonuclease digestion for various times. The deleted molecules were ligated in the presence of BamHI linkers (CCGGATCCGG) and then screened for appropriate individual deletions. Two such molecules are pKT225 and pKT229 (Figure 2). BamHI:BglII digests of these molecules are the sources of fragments 15 and 16 respectively. All three fragments were inserted into the BglII expression site of pMA91 to produce plasmids pMA91-13, pMA91-15 and pMA91-16. No new ATG potential initiation codons are generated during these constructions.

DNA sequencing

The Tyl-17 element used in the Oxford laboratory was obtained from the recombinant phage λgtKG17 (4,15). This phage contains a 7.2 Kb EcoRI fragment that contains part of the *LEU2* gene and the region centromere distal to the *LEU2* gene. Restriction fragments from this EcoRI fragment were inserted into the M13 vectors mp18 and mp19 (16) and sequenced by the dideoxy-nucleotide chain termination method of Sanger *et al.* (17). The Tyl-17 element sequenced in the Manchester laboratory came from a library of BamHI fragments constructed from the 63 u ring derivative of chromosome III (18). A 6.9 Kb BamHI fragment which maps immediately centromere-distal to the *LEU2* gene and contains the 'left' part of Tyl-17 (Figure 1a) was used as a source of fragments generated by sonication (19). The strategies used in the Oxford and Manchester laboratories are shown in Figure 3.

Nucleotide sequences were analysed in Oxford by the POLYRIB programme (A.J.Kingsman, unpublished programme). Minimal base

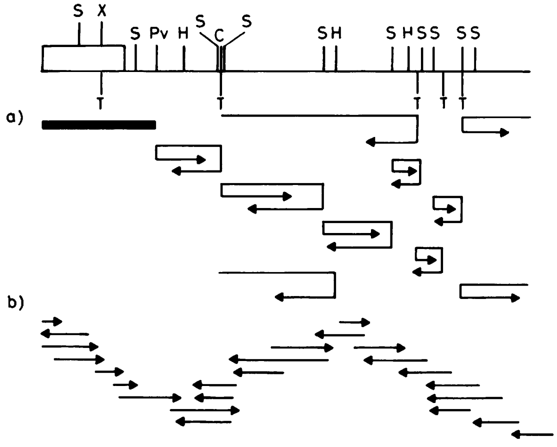

Figure 3. Sequencing strategies. The first 2.0 kb of Tyl-17 is shown. The open box is the 5' or left delta sequence. a. strategy used in the Oxford laboratory. Thick line = region sequenced and reported previously (28). Restriction fragments are marked with arrows to show the extent of the sequence obtained from each fragment. b. Strategy used in the Manchester laboratory. The extent of the sequence obtained from random fragments is marked by arrows.

change (MBC) alignment of the protein sequence was carried out using the BAZCOD programme of P.J. Artymiuk, Department of Molecular Biophysics, Oxford (personal communication). Secondary structure predictions were performed according to three methods as described by Chou and Fasman (20), Garnier *et al.* (21) and Lim (22). In Manchester variations of Staden's programmes were used to analyse the nucleotide sequence data (23,24).

Yeast extracts and protein gels

[^{35}S]Methionine-labelled protein extracts of whole yeast cells were prepared as previously described (25). Gels were run according to Laemmli (26).

Ty protein nomenclature

In the first publication describing a Ty encoded protein (8) we designated that protein p52(Tyl-15) because it was encoded by Tyl-15 and it had a mobility in polyacrylamide gels corresponding to a molecular weight of 52 Kd. However subsequent studies have shown that the mobility of this protein and the mobilities of some other Ty encoded proteins (Mellor *et al.*,manuscript in preparation) are sensitive to gel running conditions. A nomenclature based on a diagnostic gel mobility is therefore unreliable and may lead to confusion or spurious identification of 'new' proteins. We propose and use in this paper a nomenclature in which proteins are designated pl to pN, in order of their identification, followed by the name, in parentheses, of the Ty element that encodes the particular protein. Protein p52(Tyl-15) therefore becomes pl(Tyl-15) and the Tyl-17 protein identified in this paper is designated pl(Tyl-17).

RESULTS

Tyl-17 encodes a protein

Because of the heterogeneity of the Ty element family the expression of the endogenous elements can not be readily analysed. To overcome this problem we have inserted individual elements into the high efficiency expression vector pMA91 (25), this approach allowed us to identify pl(Tyl-15), the product of TYA15, on SDS-polyacrylamide gels (8). The same approach was used with Tyl-17. Plasmid pMA91 is a pBR322 derivative containing the yeast 2u plasmid origin of replication, the LEU2 selectable marker and the 5' 'promoter' and the 3' 'terminator' regions of the yeast phosphoglycerate kinase (PGK) gene. Between the 'promoter' and 'terminator' regions is a unique BglII 'expression' site (Figure 4b). Insertion of a variety of coding sequences at this site has resulted in high level expression of authentic proteins in yeast (25,27). Three fragments of Tyl-17, designated 13, 15 and 16, were constructed (see Materials and Methods). These fragments are 5' coterminal at the Ty RNA start region and extend for 5.5, 2.85 and 1.45 kb respectively (Figure 4a). The three fragments were inserted into the BglII expression site of pMA91 to produce plasmids pMA91-13, pMA91-15 and pMA91-16. Yeast strain MD40-4c was transformed with these plasmids to leucine independence to produce transformants T91-13, T91-15 and T91-16. Plasmid copy number in these transformants is about 100 and Tyl-17 homologous RNA levels are approximately 20-fold higher than the level encoded by the 35 endogenous copies of Ty (data not shown; 8).

[^{35}S]Methionine labelled proteins from T91-13, T91-15, T91-16 and a control transformant T91, containing the expression vector alone, were analysed by SDS-PAGE (Figure 5a). In extracts of T91-13, T91-15 and T91-16 a protein of about 50 Kd

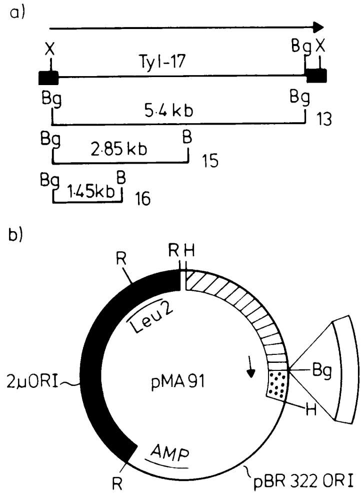

Figure 4. a. Subfragments of Tyl-17 inserted into the expression vector pMA91. Arrow marks the position of the major wild type Ty transcript. Closed boxes = delta sequences. the number beside each fragment is its designation (see text). b. The expression vector pMA91 (25). Thin line = pBR322 sequences. Closed region = 2u origin of replication, REP3 and the LEU2 gene from pJDB219. Hatched region = promoter region of the yeast PGK gene from -1500 to -1. Stippled region = transcription termination region from the PGK gene. Open box = DNA fragment to be expressed inserted at the BglII expression site. Arrow marks the direction of transcription from the PGK promoter. X = XhoI; Bg = BglII; B = BamHI; H = HindIII; R = EcoRI.

is overproduced compared with the extract of T91. The region of Tyl-17 required for expression of this protein is contained within the 1.45 kb fragment 16 as equal amounts of the same protein are produced by all three transformants. The protein is designated pl(Tyl-17) as the first identified protein associated with Tyl-17 (see Materials and Methods for nomenclature convention).

Nucleotide sequence analysis.

pl(Tyl-17) is encoded by the region of Tyl-17 that corresponds to TYA15 in Tyl-15. In order to determine the genetic organisation of this region of Tyl-17 and the primary sequence of pl(Tyl-17) the nucleotide sequence of the first 2.0 kb of Tyl-17 was obtained. The sequence was determined independently by both the Oxford and Manchester laboratories (see Materials and Methods) and both sets of results were identical. The sequence of the first 1630 nucleotides of the element compared with the same region of Tyl-15 is shown in Figure 6. The two sequences have been aligned to match the alignment of the amino acid sequences of the proteins encoded by these regions (see later; Figure 7). The delta sequences of these two elements are very similar and have been compared previously (28). The first striking feature of the Tyl-17 sequence is the presence of an open reading frame (ORF) extending from nucleotide 1 (Figure 6) to nucleotide 1314 that would encode a protein of 50,897 Kd. This correlates with the

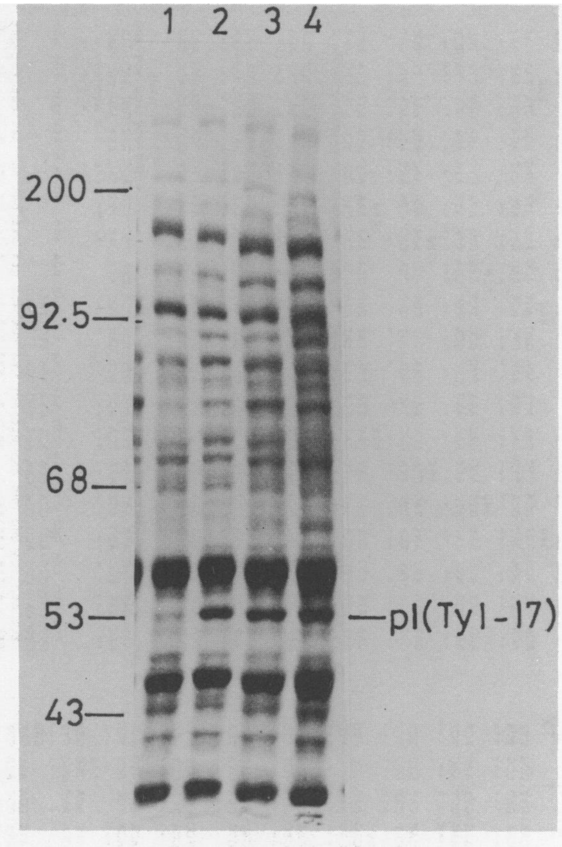

Figure 5. SDS-polyacrylamide gel analysis of ^{35}S-methionine labelled extracts of T91 (lane 1), T91-13 (lane 2), T91-15 (lane 3) and T91-16 (lane 4). The positions and sizes of molecular weight markers are shown on the left. The position of pl(Tyl-17) (see text) is shown on the right.

overproduction of pl(Tyl-17) in T91-13, T91-15 and T91-16 and we would conclude that this ORF encodes pl(Tyl-17). There are no other major ORF's in this region. The position of the coding sequence for pl(Tyl-17) in Tyl-17 matches that of TYA15 in Tyl-15. The coding sequence is therefore designated TYA17. TYA15 and TYA17 are about 64% homologous in the relative positions shown in Figure 6. However the homology is dispersed with the longest stretch of continuous homology being the 26 residues from 54 to 79. The dispersed but discernable homology suggests that the difference between the two genes arose primarily by the

```
TGTTGGAATA AAAATCAACT ATCATCTACT AACTAGTATT TACGTTACTA GTATATTATC ATATACGGTG
TGTTGGAATA GAAATCAACT ATCATCTACT AACTAGTATT TACATTACTA GTATATTATC ATATACGGCT
==========  ========= ========== ========== === ====== ========== ========

---TTAGAAG ATGACGCAAA TGATGAGAAA TAGTCATCTA AATTAGTGGA AGCTGAAACG CAAGGATTGA
GGCTTAGAAG ATGACGCAAA TGATGAGAAA TAGTCATCTA AATTAGTGGA AGCTGAAACG CAAGGATTGA
   ======= ========== ========== ========== ========== ========== ==========

TAATGTAATA GGATCAATGA AATATTAACA TATAAAATG- -ATGAATAAT ATTTATAGAA TTGTGTAGAA
TAATGTAATA GGATCAATGA AATATAAACA TATAAAACGG AATGAATAAT CGTAATATTA GTATGTAGAA
========== ========== ===== ==== =======     =========   = ===  =  = =======

TTGCAGATTC CCTTTTATGG ATTCCTAAAT CCCTCGAGGA GAACTTCTAG TATAT-CTAC ATACCTAATA
ATATAGATTC CATTTTGAGG ATTCCTATAT CCCTCGAGGA GAACTTCTAG TATATTCTGT ATACCTAATA
 =  ====== = ====  == ======= == ========== ========== ===== ==   ==========

                    +1                       20                        40 v
TTATTGCCTT  -ATAAAAA ATG GAA TCC CAA CAA TTA CAT CAA AAT CCA CGT TCT CTT CAT
TTATAGCCTT  TATCAACA ATG GAA TCC CAA CAA TTA TCT CAA CAT TCA CCC ATT TCT CAT
==== =====   == == = === === === === === ===   = ===  ==  == =     =   = ===
                                                                          ^
                      60                         80
GGT AGC GCC TAT GCT TCG GTT ACT TCT AAG GAA GTC CCA TCA AAT CAA GAT CCG TTA
GGT AGC GCC TGT GCT TCG GTT ACT TCT AAG GAA GTC CAC ACA AAT CAA GAT CCG TTA
=== === === = = === === === === === === === === =    == === === === === ===

                         120                        140
GCC GTT TCA GCT TCC AAT TTA CCG GAA TTT GAT AGA GAT TCC ACT AAG GTT AAT TCT
GAC GTT TCA GCT TCC AAA ACA GAA GAA TGT GAG AAG GCT TCC ACT AAG GCT AAC TCT
= = === === === === ==    =     === = = ==  =   = = === === === = = ==  ===

   160                       180                         200
CAA CAA GAG ACA ACA CCT GGG ACA TCA GCT GTT CCA GAG AAC CAT CAT CAT GTC TCT
CAA CAG ACA ACA ACA CCT GCT TCA TCA GCT GTT CCA GAG AAC CCC CAT CAT GCC TCT
=== ==      === === === =    == === === === === === === =   === === = = ===

       220                       240                         260
CCT CAA CCT GCT TCA GTA CCA CCT CCA CAG AAT GGA CAG TAC CAA CAG CAC GGC ATG
CCT CAA ACT GCT CAG TCA CAT TCA CCA CAG AAT GGG CCG TAC CCA CAG CAG TGC ATG
=== ===  == ===       = =       === === === ==  = = === = = === ==   == ===

           280                       300                         320
ATG ACC CCA AAC AAA GCT ATG GCC TCT AAC TGG GCA CAT TAC CAA CAA CCG TCT ATG
ATG ACC CAA AAC CAA GCC AAT CCA TCT GGT TGG TCA TTT TAC GGA CAC CCA TCT ATG
=== === = = ===  == ==  =    =  ===     ===  ==   = ===   = ==  ==  === ===

               340                       360                         380
ATG ACG TGT TCA CAT TAT CAA ACG TCA CCT GCG TAT TAT CAA CCG GAC CCA CAC TAT
ATT CCG TAT ACA CCT TAT CAA ATG TCG CCT ATG TAC TTT CCA CCT GGG CCA CAA TCA
==   == = =  == = = === === = = ==  ===   = ==  = = = = ==  =   === ==  =

                   400                       420                         440
CCG TTG CCA CAG TAT                     ATC CCA CCA CTG AGT ACT TCC TCA CCT GAT
CAG TTT CCG CAG TAT CCA TCA TCA GTT GGA ACG CCT CTG AGG ACT CCA TCA CCT GAG
= = ==  ==  === ===                      =  ==  === ==  ===  =  === === ==

                       460                           480
CCA ATC GAT TCA CAG AAT CAA CAC TCT GAA GTA CCT CAA GCT GAG ACA AAG GTG AGA
TCA GGT AAT ACA TTT ACT GAT TCA TCC TCA GCG GAC TCT GAT ATG ACA TCC ACT AAA
 ==      ==  ==     = =  =      ==    = =           = =   = ===         = =

                500                        520                        540
AAT AAC GTC TTA CCA CCA CAC ACT TTA ACA TCA GAA GAA AAC TTT TCT ACA TGG GTT
AAA TAT GTC AGA CCA CCA CCA ATG TTA ACC TCA CCT AAT GAC TTT CCA AAT TGG GTT
==   =  ===   = === === =   =   === ==  ===      =   == ===  =  =   === ===

                    560                        580                        600
AAA TTT TAC ATC AGA TTT TTG AAG AAC TCT AAT CTC GGT GAC ATC ATT CCA AAT GAC
AAA ACA TAC ATC AAA TTT TTA CAA AAC TCG AAT CTC GGT GGT ATT ATT CCG ACA GTA
===     === === = = === ==   =  === ==  === === === =   ==  === ==  =   =

                        620                         640
CAG GGT GAA ATC AAA AGA CAA ATG ACT TAT GAA GAA CAT GCG TAT ATA TAC AAT ACC
AAC GGA AAA CCC GTA CGT CAG ATC ACT GAT GAT GAA CTC ACC TTC TTG TAT AAC ACT
 =  ==   ==   =   =  =  ==  ==  ===  == ==  === =    =  =    =  ==  ==  ==

660                         680                         700
TTC CAA GCA TTT GCC CCA TTT CAT TTA TTG CCA ACA TGG GTA AAA CAA ATT TTA GAA
TTT CAA ATA TTT GCT CCC TCT CAA TTC CTA CCT ACC TGG GTC AAA GAC ATC CTA TCC
==  ===   = === ==  ==  = = ==  ==   =  ==  ==  === ==  ===  =  ==   ==

     720                        740                         760
ATT AAT TAT GCT GAC ATC CTT ACA GTC CTT TGT AAA AGT GTG TCC AAA ATG CAA ACT
GTT GAT TAT ACG GAT ATC ATG AAA ATT CTT TCC AAA AGT ATT GAA AAA ATG CAA TCT
 ==  == ===  =  ==  ===  =  = =  =  === =   === ===  =      === === ===  ==

         780                         800                        820
AAC AAT CAA GAA TTA AAG GAT TGG ATA GCT CTT GCC AAC CTT GAG TAC GAC GGA AGT
GAT ACC CAA GAG GCA AAC GAC ATT GTG ACC CTG GCA AAT TTG CAA TAT AAT GGC AGT
 =  =   === ==    = ==  ==       =   =  ==  ==  ==   =   =  ==   =  ==  ===

             840                        860                        880
ACA TCT GCT GAT ACA TTT GAA ATT ACA GTC AGT ACG ATC ATT CAG AGG CTA AAA GAA
ACA CCT GCA GAT GCA TTT GAA ACA AAA GTC ACA AAC ATT ATC GAC AGA CTG AAC AAT
===  == ==  ===  == === === =   = = === =   =   ==  ==   =  ==  ==  ==   =

                 900                        920                        940
AAC AAT ATC AAT GTT AGC GAC AGA TTG GCC TGT CAA CTA ATA CTT AAA GGT CTA TCC
AAT GGC ATT CAT ATC AAT AAC AAG GTC GCA TGC CAA TTA ATT ATG AGA GGT CTA TCT
==      ==   ==  =  =    ==  =    =  ==  ==  ===  == ==   =  = = === === ==

                     960                        980
GGT GAC TTC AAA TAC CTA CGT AAT CAA TAT CGT ACC AAA ACG AAC ATG AAA CTT TCC
GGC GAA TAT AAA TTT TTA CGC TAC ACA CGT CAT CGA CAT CTA AAT ATG ACA GTC GCT
==  ==  =   === =    == ==   =    =   = = =      =      ==  === = =  =   =

                         1020                       1040
CAA TTA TTC GCT GAA ATT CAG TTA ATA TAT GAC GAA AAT AAA ATC ATG AAT CTA AAT
GAA CTG TTC TTA GAT ATC CAT GCT ATT TAT GAA GAA CAA CAG GGA TCG AGA AAC AGT
 ==  =  ===     ==  ==  ==      ==  === ==  ===  =   =        = =       = =

   1060                      1080                       1100
AAA CCG TCC CAA TAC AAA CAA CAC AGC GAA TAC AAA AAT GTT TCT CGC ACA TCT CCA
AAA CCT AAT TAC AGG AGA AAT CCG AGT GAT GAG AAG AAT GAT TCT CGC AGC TAT ACG
=== ==       =      = =  =  =   ==  ==   =  ==  === = = === === =   = =  =

       1120                      1140                       1160
AAC ACG ACT AAC ACA AAG GTT ACA ACT CGT AAT TAT CAG AGA ACA AAT AGT TCA AAA
AAT ACA ACC AAA CCC AAA GTT ATA GCT CGG AAT CCT CAA AAA ACA AAT AAT TCG AAA
==  ==  ==  ==   =  ==  === = =  == ==  ===   = ==  = = === === = = ==  ===

           1180                      1200                       1220
CCA AGA GCA GCA AAA GCT CAC AAT ATT GCT ACA TCT AGT AAA TTC TCA AGG GTG AAC
TCG AAA ACA GCC AGG GCT CAC AAT GTA TCC ACA TCT AAT AAC TCT CCC AGC ACG GAC
 =  = =  == ==  =   === === ===  =   =  === === = = ==  =    =  ==    =  ==

                                                                   1280
               1240                      1260            gcg atg aca acg aac
AAT GAT CAC ATT AAT GAA TCA ACC GTT TCA TCA CAA TAC TTA AGC GAT GAC AAC GAA
AAC GAT TCC ATC AGT AAA TCA ACT ACT GAA CCG ATT CAA TTG AAC AAT AAG CAC GAC
==  ===  == ==  = =  == === ==    =   =  =       =  ==  =aca ata agc acg acc
                                                          =  ==  =   =   ==

                   1300
 tta gtc tta ggc cag caa cag aaa gaa tct aag cca aca cac aca ata gac tcg aat
CTT AGT CTT AGG CCA GCA ACA GAA AGA ATC TAA.
CTT CAT CTT AGG CCA GAA ACT TAC TGA.
 ttc atc tta ggc cag aaa ctt act gaa tct aca gta aat cat act aat cat tct gat
 ==   == === === ===  == =   =   === === =     = =   ==  ==  =    =  ==   ==
```

Figure 6. The sequence of the first 1630 nucleotides of Tyl-17. Upper sequence = Tyl-17; Lower sequence Tyl-15. The sequences are aligned according to the alignment of the protein sequences in Figure 7. The arrow heads mark the boundary between the delta and internal epsilon regions. The open reading frame corresponding to TYA encoding the pl proteins is represented in triplets. The beginning of the overlapping TYB open reading frame is shown in lower case. An '=' sign marks identical residues in the two elements. S = Sau3a; X = XhoI; T = TaqI, Pv = PvuII; H = HaeIII; C = ClaI.

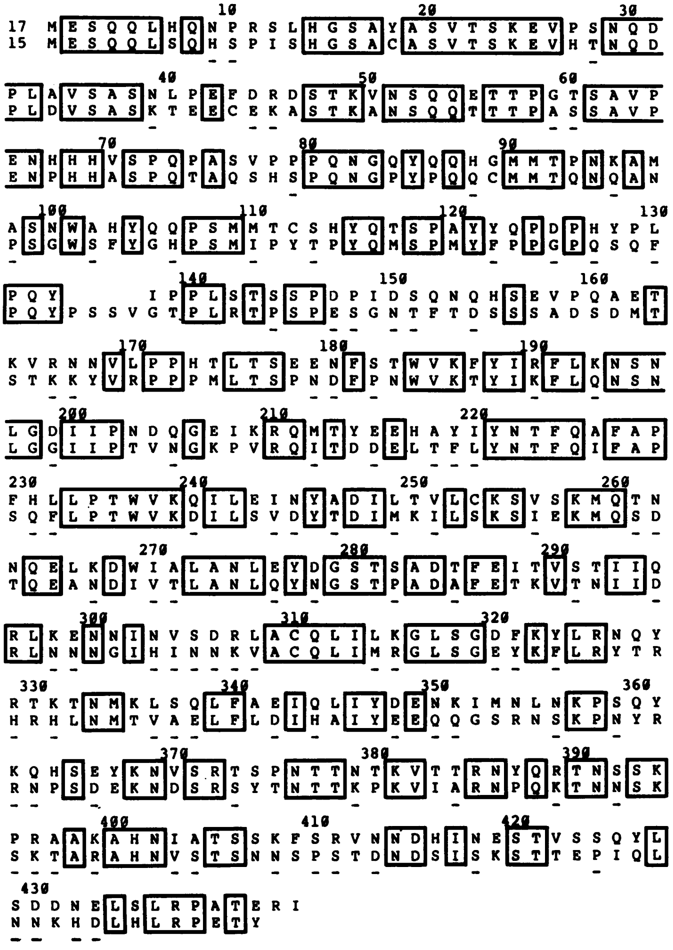
10 20 30
17 M E S Q Q L H Q N P R S L H G S A Y A S V T S K E V P S N Q D
15 M E S Q Q L S Q H S P I S H G S A C A S V T S K E V H T N Q D
40 50 60
P L A V S A S N L P E F D R D S T K V N S Q Q E T T P G T S A V P
P L D V S A S K T E E C E K A S T K A N S Q Q T T T P A S S A V P
70 80 90
E N H H H V S P Q P A S V P P P Q N G Q Y Q Q H G M M T P N K A M
E N P H H A S P Q T A Q S H S P Q N G P Y P Q Q C M M T Q N Q A N
100 110 120 130
A S N W A H Y Q Q P S M M T C S H Y Q T S P A Y Y Q P D P H Y P L
P S G W S F Y G H P S M I P Y T P Y Q M S P M Y F P P G P Q S Q F
140 150 160
P Q Y I P P L S T S S P D P I D S Q N Q H S E V P Q A E T
P Q Y P S S V G T P L R T P S P E S G N T F T D S S S A D S D M T
170 180 190
K V R N N V L P P H T L T S E E N F S T W V K F Y I R F L K N S N
S T K K Y V R P P P M L T S P N D F P N W V K T Y I K F L Q N S N
200 210 220
L G D I I P N D Q G E I K R Q M T Y E E H A Y I Y N T F Q A F A P
L G G I I P T V N G K P V R Q I T D D E L T F L Y N T F Q I F A P
230 240 250 260
F H L L P T W V K Q I L E I N Y A D I L T V L C K S V S K M Q T N
S Q F L P T W V K D I L S V D Y T D I M K I L S K S I E K M Q S D
270 280 290
N Q E L K D W I A L A N L E Y D G S T S A D T F E I T V S T I I Q
T Q E A N D I V T L A N L Q Y N G S T P A D A F E T K V T N I I D
300 310 320
R L K E N N I N V S D R L A C Q L I L K G L S G D F K Y L R N Q Y
R L N N N G I H I N N K V A C Q L I M R G L S G E Y K F L R Y T R
330 340 350 360
R T K T N M K L S Q L F A E I Q L I Y D E N K I M N L N K P S Q Y
H R H L N M T V A E L F L D I H A I Y E E Q Q G S R N S K P N Y R
370 380 390
K Q H S E Y K N V S R T S P N T T N T K V T T R N Y Q R T N S S K
R N P S D E K N D S R S Y T N T T K P K V I A R N P Q K T N N S K
400 410 420
P R A A K A H N I A T S S K F S R V N N D H I N E S T V S S Q Y L
S K T A R A H N V S T S N N S P S T D N D S I S K S T T E P I Q L
430
S D D N E L S L R P A T E R I
N N K H D L H L R P E T Y

Figure 7. Deduced protein sequence of pl(Tyl-17). Upper sequence = pl(Tyl-17); lower sequence = pl(Tyl-15). See text for alignment. Identical residues are boxed; a '-' sign marks a change that has a positive score on a Dayhoff log odds matrix (29).

accumulation of point mutations, rather than by block substitution resulting from events such as transposition or conversion. Remarkably the organisation of the ORF's has been conserved despite this considerable variation. Both genes are about the same size and begin and end in approximately the same relative positions and the relationships of both genes to a second overlapping ORF (shown in lower case in Figure 6) are the

```
        10        20        30        40        50        60        70        80        90       100
         .         .         .         .         .         .         .         .         .         .
AAAAAA--TTTT-------AAAAAAATTTT----------AAAAAA------bbbbbTTT-TTTTTTTT--TTTTT----TTTT--bbbbbbb---AAAA
AAAAAATTTTTTTT-----aaaaaaa----------AAAAAAAAAAAAAA--BBBBB--------TTTT-TTTTTTTTT-TTTT--BBBBBBB---TTTT
=====   ====  ============    ======    ======    =======   =    ===== =====   =================

       110       120       130       140       150       160       170       180       190       200
         .         .         .         .         .         .         .         .         .         .
AAA---TTBBBBB-----BBBBBTTTTT-----    -TTTTT-TTTT--TTTT----AAAAAAAA-----------------BBBBBBBB--TTTTTT-
------TTBBBBB-----BBBBBTTTTT----TTTTTTTTTT--TTTT-----TTTT-TTTT----BBBB-------TTT----BBBBBBBBBTTTTT--
   =============================      ===  =======   =   =            =======   === =======  ===== =

       210       220       230       240       250       260       270       280       290       300
         .         .         .         .         .         .         .         .         .         .
TTTT--AAAAAAAAAAAABBBBBB-----------BBBBBBBBBBBBBBBB--AAAAAAA-----AAAAAAAAAATTTTT-TTTBBBBBBBBBBBBB-TT
--TTTT-----------BBBBBBBBBBTTTT------BBBBBBBBBBBBBB-AAAAAAAAAAAAAABBBBBBBB-TTTT------BBBBBBBBB---TTT
  ==              ======       ====  =============== =======     =         ==== =    =========    ==

       310       320       330       340       350       360       370       380       390       400
         .         .         .         .         .         .         .         .         .         .
TTTT--BBBBBBBBB-TTTTBBBBB---------AAAAAAAAAA------BBBBB-TTTT----TTTTTTT-TTTT---BBBBBBTTTTTTT---AAAAA
TBBBBBBBBBBBBB--TTTTBBBBBB----------AAAAAAAAAAAAA-TTTTTTTTT-----TTTTTTTTTTTTT-AAAAAAA---------------
=     ========  ========= ========  ========     =      === =========== ==== =              ===

       410       420       430       440
         .         .         .         .
----TTTT----TTTT--TTTTTT---TTT----AAAAAA--
----TTTTTTTT-----TTTT---------------TTTT-
========        = ===   ===   ====
```

Figure 8. Secondary structure of p1 proteins. 'A' = alpha helix, 'T' = turn, 'B' = beta sheet, '---' = unstructured, '=' = structural identity. Secondary structure designations are marked only where at least two of the three predictive methods (see Materials and Methods) agree. Exceptions, where only one method predicted structure and the others predicted no structure, are represented by lower case letters. The sequence of p1(Ty1-17) is the upper line and p1(Ty1-15) is the lower line. Numbering corresponds to amino acid residues aligned as in Figure 7.

same. This second ORF is the TYB gene that has been shown to extend to the end of a Ty1-15-like element (1). When both nucleotide sequences are aligned, as in Figure 6, to maximise protein sequence homology, TYB15 and TYB17 start at the same place. TYB15 overlaps TYA15 by 38 nucleotides and TYB17 overlaps TYA17 by 44 nucleotides. In both cases TYA and TYB are out of phase with respect to translational reading frame. The amino acid sequences of TYA15 and TYA17 are compared in Figure 7. The sequences are aligned by a minimal base change analysis. This analysis suggests that maximum alignment is achieved if a deletion/insertion of four residues occurs at residue 134 to 137. On the basis of this alignment, the two proteins display 48% sequence identity with 52% of the remaining substitutions being relatively common changes (29). This level of similarity is comparable with other structurally related protein families. For example, there is 60-70% sequence identity observed in the structurally conserved beta-barrel framework regions of different immunoglobulin Fab fragments (P.de la Paz, unpublished results) The Ty proteins are therefore broadly similar. The proteins are of a similar size, pI (7.8) and amino acid composition and their coding sequences have a relatively low Bennetzen and Hall (30) codon bias index (0.31) indicating that TYA17 and TYA15 may not be expressed efficiently. In order to extend this comparison both amino acid sequences were subjected to a three way (see Materials and Methods) secondary structure prediction calculation (Figure 8). The structural predictions suggest that: 54% of the residues in the two proteins can be assigned secondary structure; secondary structure is conserved in about 60% of the residues; the proteins belong to the alpha/beta class. Main areas of difference occur at residues 266-276 and 380-385 but otherwise the distribution of secondary structure features is remarkably similar.

DISCUSSION

We have set out to establish whether two Ty elements, Tyl-15 and Tyl-17, which show the maximum differences observed to date (4) share any features of their gene products and expression strategies. In order to assess any eukaryotic gene expression strategy it is essential to analyse gene products as well as nucleic acid sequence because DNA sequence information alone cannot predict, for example, splicing or frameshifting. This is particularly pertinent to Ty elements as a complex frameshifting event has already been demonstrated as part of the expression strategy of the TYB region of Tyl-15 (2). In order to analyse Tyl-17 gene products we have inserted various parts of the element into a high efficiency yeast expression vector. We have shown that a 1.45 kb fragment (fragment 16) containing part of the 5' delta and part of the epsilon region of Tyl-17 directs the production of a 50 Kd protein, pl(Tyl-17). This fragment spans a region of the element which has no apparent homology to Tyl-15 in electronmicroscopic heteroduplex analyses (4) yet Tyl-15 also encodes a 50 Kd protein, pl(Tyl-15), in this region (8). Nucleotide sequence data revealed single long open reading frames in both Tyl-15 (2) and Tyl-17 that could encode these proteins suggesting that the pl proteins are the simple primary translation products of these regions. A comparison of the nucleotide sequences reveals many base changes which account for the lack of heteroduplex formation (4). The most striking feature is, however, the high degree of structural homology when the amino acid sequences of the pl proteins are compared. The two proteins are homologous at 48% of the amino acid positions and there are many conservative changes. The similarity is further emphasised by the distribution of secondary structural features.

The finding that the structure of the pl proteins of Ty elements has been conserved, despite considerable variation in their DNA sequences, strongly suggests that selection has acted on them. These observations provide, therefore, the first indication that the proteins encoded by these elements have function. This argument therefore dispels the idea that Ty elements are 'junk' DNA and raises questions as to the nature of the functions of Ty proteins and the level at which selection for these functions acts. Whatever the function of the pl proteins, selection for their structural conservation could operate at one or both of two levels. The pl proteins may confer on the organism a phenotype that is subject to selection. Massive overproduction of either pl(Tyl-15) or pl(Tyl-17) alone does not result in any apparent change in phenotype (unpublished data). However, in a limited laboratory environment the 'correct' phenotype may not have been tested. High level expression of the TYB regions is in fact toxic (unpublished data). A second possiblity is that Ty has no effect on its 'host' and that selection for Ty-encoded functions operates at the genomic level within the broad terms of the concept of 'selfish DNA' (31,32). The Ty element would only 'require' preservation of its transposition functions to maintain a controlled exploitation of the yeast genome and it would be this 'requirement' that would be the selective force for structural, and thereby functional, conservation of its proteins.

Given the conservation observed at the protein level it is somewhat curious that the Ty elements show such marked DNA sequence heterogeneity. The 35 copies of the element in the haploid genome of most laboratory strains of yeast have been shown to interact readily by gene conversion (6), a process that can lead to homogeneity in the absence of a counteracting process that generates diversity at a relatively high rate. In considering mechanisms that may generate diversity, and given the similarities of retroviruses and Ty elements (1,2,3) it may be useful to think of the Ty element as an endogenous retrovirus that cannot be transmitted horizontally. Its dissemination, therefore, is either vertical or by spreading within its host genome by transposition. This would be achieved via an RNA intermediate that would be reverse transcribed prior to integration. Reverse transcriptase is a highly error prone enzyme (33) generating mistakes at frequencies as high as 1 per 500 nucleotides and, therefore, could be the major factor in generating diversity amongst Ty elements. However, this argument is not entirely satisfactory as it does not explain why the copia-like elements are highly conserved even though one would expect these elements to move via reverse transcriptase nor does it provide a mechanism for conservation of structure when considerable variation is maintained within a single sequence family. It may be more appropriate to think of the class I and class II elements as sub-families.

Clearly the conservation of pl proteins implies that they have a key role in Ty biology. At present that role is unknown but the striking similarities between Ty elements and retroviruses suggests that a search for nucleic acid packaging and RNA-dependent and DNA-dependent polymerase activities might be appropriate.

ACKNOWLEDGEMENTS

We would like to thank Sir David Phillips for advice and encouragement throughout the work and we are grateful to Philip Farabaugh and Horst Feldmann for communicating preliminary DNA sequence information prior to publication. The Manchester group is also grateful to John Guest, Ian Duckenfield and Martin Bishop for advice on sequence assembly and Richard Waring and Wayne Davies for many helpful discussions. The work was supported in Oxford by the MRC (G8317537CB) and the Nuffield Foundation and in Manchester by the SERC (GR/C34656) and by a NATO Collaborative Research Award to SGO and Dr C.S.Newlon, Department of Zoology, University of Iowa, Iowa, U.S.A. We are particularly indebted to Dr Newlon for the provision of clone A5C and for helpful comments and support.

*To whom correspondence should be addressed

REFERENCES

1. Hauber, J., Nelbock-Hockstetter, P. and Feldmann, H. (1985) Nucl. Acids Res. **13**, 2745-2758.
2. Mellor, J., Fulton, A.M., Dobson, M.J., Wilson, W., Kingsman, S.M., and Kingsman, A.J. (1985) Nature **313**, 243-246.
3. Varmus, H.E. (1983) in Shapiro, J.A. (Ed) Mobile Genetic Elements Academic Press, N.Y., pp 411-503.
4. Kingsman, A.J., Gimlich, R.L., Clarke, L., Chinault, A.C.

and Carbon, J. (1981) J.Mol.Biol. **145**, 619-632.

5. Rubin, G.M. (1983) in Shapiro, J.A. (Ed) Mobile Genetic Elements Academic Press, N.Y., pp 329-361.

6. Roeder, G.S. and Fink, G.R. (1982) Proc. Natl. Acad. Sci. USA **79**, 5621-5625.

7. Elder, R.T., Loh, E.Y. and Davis, R.W. (1983) Proc. Natl. Acad. Sci. USA **80**, 2432-2436.

8. Dobson, M.J., Mellor, J., Fulton, A.M., Roberts, N.J., Bowen, B.A., Kingsman, S.M. and Kingsman, A.J. (1984) EMBO.J. **3**, 1115-1119.

9. Miller, J.H. (1972) Experiments in molecular genetics. Cold Spring Harbor Laboratory, N.Y.

10. Hawthorne, D.C. and Mortimer, R.K. (1960) Genetics **45**, 1085-1110.

11. Chinault, A.C. and Carbon, J.A. (1979) Gene 5, 111-126.

12. Holmes, D.S. and Quigley, M. (1981) Analyt. Biochem. **114**, 193-197.

13. Hinnen, A., Hicks, J.B. and Fink, G.R. (1978) Proc. Natl. Acad. Sci. USA **75**, 1929-1933.

14. Tabak, H.F. and Flavell, R.A. (1978) Nucl. Acids. Res. **5**, 2321-2332.

15. Dobson, M.J., Kingsman, S.M. and Kingsman, A.J. (1981) Gene **16**, 133-139.

16. Norrander, J., Kempe, T. and Messing, J. (1983) Gene **26**, 101-106.

17. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA **74**, 5463-5467.

18. Newlon, C.S., Devenish, R.J. and Lipschitz, L.R. (1982) Rec. Adv. Yeast Mol. Biol. **1**, 52-68.

19. Deininger, P.L. (1983) Anal. Biochem. **129**, 216-223.

20. Chou, P.Y. and Fasman, G.D. (1974) Biochemistry **13**, 222-244.

21. Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) J. Mol. Biol. **120**, 97-120.

22. Lim, V.I. (1974) J. Mol. Biol. **88**, 873-894.

23. Staden, R. (1981) Nucl. Acids Res. **8**, 3673-3694.

24. Staden, R. and McLachlan, A.D. (1982) Nucl. Acids. Res. **10**, 141-149.

25. Mellor, J., Dobson, M.J., Roberts, N.A., Tuite, M.F., Emtage, J.S., White, S., Lowe, P.A., Patel, T., Kingsman, A.J. and Kingsman, S.M. (1983) Gene **24**, 1-14.

26. Laemmli, U. K. (1970) Nature **227**, 680-685.

27. Kingsman, S.M. and Kingsman, A.J. (1983) in Burke, D.C. and Morris, A.G. (Eds) Interferons: From Molecular Biology to Clinical Application. Society for General Microbiology Symposium 35. C.U.P. pp 212-254.

28. Bowen, B.A., Fulton, A.M., Tuite, M.F., Kingsman, S.M. and Kingsman, A.J. (1984) Nucl. Acids Res. **12**, 1627-1640.

29. Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978) in Atlas of Protein Sequences**5**, suppl. 3, 345-351.

30. Bennetzen, J.L. and Hall, B.D. (1983) J. Biol. Chem. **257**, 3026-3031.

31. Orgel, L.E. and Crick, F.H.C. (1980) Nature **284**, 604-607.

32. Doolittle, W.F. and Sapienza, C. (1980) Nature **284**, 601-603.

33. Mizutani, S. and Temin, H.M. (1976) Biochemistry **15**, 1510-1516.