Nucleic Acids Research, 1992, Vol. 20, No. 24 6621–6629

# *Euplotes crassus* has genes encoding telomere-binding proteins and telomere-binding protein homologs

Wenlan Wang, Rose Skopp, Margaret Scofield[1] and Carolyn Price*
Department of Chemistry, University of Nebraska, Lincoln, NE 68588 and [1]Department of Pharmacology, Creighton University Medical School, Omaha, NE 68178, USA

Received August 31, 1992; Revised and Accepted October 19, 1992 GenBank accession nos M96818, M96819

## ABSTRACT

**We have identified two 1.6 kb macronuclear DNA molecules from *Euplotes crassus* that hybridize to the α subunit of the *Oxytricha* telomere protein. We have shown that one of these molecules encodes the 51 kDa *Euplotes* telomere protein while the other appears to encode a homolog of the telomere protein. Although this homolog clearly differs in sequence from the *Euplotes* telomere protein, the two proteins share extensive amino acid sequence identity with each other and with the α subunit of the *Oxytricha* telomere protein. In all three proteins 35–36% of the amino acids are identical, while 54–56% are similar. The most extended regions of sequence conservation map within the N-terminal section; this section has been shown to comprise the DNA-binding domain in the *Euplotes* telomere protein. Our findings suggest that some of the conserved amino acids may be involved in DNA recognition and binding. The gene encoding the telomere protein homolog contains two introns; one of these introns is only 24 bp in length. This is the smallest mRNA intron reported to date.**

## INTRODUCTION

Telomeres are the structures found at the termini of linear eukaryotic chromosomes. They have several vital functions including protection of chromosome ends from degradation or fusion, and assisting in the complete replication of the most terminal DNA sequences (reviewed in 1–3). They can also influence the architecture of the nucleus (4), possibly via their association with the nuclear matrix (4). Telomeres exist as complexes of DNA and protein that differ both in structure and composition from nucleosomes (6–11).

The DNA component of telomeres consists of tandem repeats of a short 5–9 base pair sequence. Although the exact sequence of the telomeric DNA varies from species to species, clusters of 3 or more G residues are usually found on the strand that runs towards the 3′ end of the DNA (1–3, 12). For example, the ciliates *Euplotes* and *Oxytricha* both have telomeric DNA that consists of the sequence $C_4A_4.T_4G_4$ (13). *Euplotes* has 28 bp of this sequence in addition to an extra 14 base $T_4G_4T_4G_2$ extension on the end of the 3′ strand. In *Oxytricha* the double-stranded region of the telomere contains only 20 bp of the $C_4A_4.T_4G_4$ sequence, while the extension on the 3′ strand is 16 bases in length and has the sequence $T_4G_4T_4G_4$. The extension on the 3′ strand is a characteristic of telomeres from a number of organisms and may be a feature of all telomeres (14).

Telomere-binding proteins have been isolated from *Physarum*, yeast, *Oxytricha*, and *Euplotes* (1,15–23). Those from *Physarum* (PPT) and yeast (TBPα and RAP1) bind to internal stretches of telomeric DNA. PPT is thought to coat the entire double-stranded ($T_2AG_3$) telomeric DNA (15) while TBP α may bind at the junction between the $G_{1-3}T$ repeats and the subtelomeric X sequence (16). RAP1 recognizes a consensus sequence which occurs in upstream activator, silencer and telomeric DNA sequences (1). Abnormal expression of RAP1 results in chromosome instability and deregulation of telomere length control (17–20).

The telomere proteins from *Oxytricha nova* and *Euplotes crassus* bind to the extreme end of telomeric DNA rather than to internal sequences (21–23). *In vivo* they are thought to form a protective cap over the ends of each macronuclear DNA molecule. However, they may also participate in telomere length regulation during DNA replication (12,22,23). Both the *Oxytricha* and *Euplotes* telomere proteins bind very specifically to the $T_4G_4$-containing extension on the 3′ strand and protect this region of the telomere from nuclease digestion and chemical modification (22,23). Although both proteins bind telomeric DNA non-covalently, the association is tenacious as neither protein is dissociated by high salt (2 M NaCl or 6 M CsCl) and removal of the DNA has only been achieved by denaturing the proteins or digesting away the DNA (23,24).

Characterization of the *Oxytricha* and *Euplotes* telomere proteins has been facilitated by their relative abundance. In hypotrichous ciliates such as *Oxytricha* and *Euplotes*, the macronuclear genome is composed of $\sim 2\times10^7$ separate gene-sized DNA molecules (reviewed in 25). These molecules contain all the regulatory elements that are required for gene expression and DNA replication. As each molecule has telomeres on either end, a macronucleus contains $\sim 4\times10^7$ telomeres and a correspondingly large amount of telomere-binding protein.

* To whom correspondence should be addressed

The *Oxytricha* telomere protein is a 97 kDa heterodimer composed of a 56 kDa $\alpha$ and a 41 kDa $\beta$ subunit (21,22,24,26). The two subunits of the native protein are very tightly associated and it has not been possible to separate them without destroying the DNA-binding activity of the protein (24). Examination of the DNA-binding properties of the individual subunits only became possible when the genes encoding each subunit were cloned and expressed in *E.coli* (26,27). *In vitro* studies with the expressed subunits showed that the $\alpha$ subunit is the dominant DNA-binding moiety. However, the $\beta$ subunit is required to achieve full *in vivo* binding activity (26). The genes encoding the $\alpha$ and $\beta$ subunits of the *Oxytricha* protein have been used to clone the two equivalent genes from a closely related ciliate *Stylonychia mytilis* (28). Although the *Stylonychia* telomere protein has not yet been isolated, it is likely to be very similar to the *Oxytricha* telomere protein as the peptides encoded by the *Stylonychia* genes are 79% and 77% identical to the $\alpha$ and $\beta$ subunits of the *Oxytricha* protein.

The *Euplotes* telomere protein differs from the *Oxytricha* protein in that it has been isolated as a single subunit of 51 kDa rather than as a heterodimer (23). It appears that the *Euplotes* protein does not require a second subunit to bind telomeric DNA in a sequence-specific and salt-stable manner (23,29). In fact, the DNA-binding domain has been mapped to a specific region of the protein (29). Trypsin digestion releases a ~35 kDa protease resistant peptide from the N-terminus, this peptide retains most of the DNA-binding characteristics of the native protein. We have now cloned and sequenced the gene that encodes the *Euplotes* 51 kDa telomere protein (previously referred to as the 50 kDa protein (23)). In this paper we present evidence that the *Euplotes* telomere protein shares sequence identity with the $\alpha$ subunit of the *Oxytricha* and *Stylonychia* telomere proteins. We have also cloned a gene that appears to encode a homolog of the *Euplotes* telomere protein. This protein shares striking sequence identity with both the *Euplotes* telomere protein and the $\alpha$ subunit of the *Oxytricha* and *Stylonychia* telomere proteins. The regions that are most highly conserved between the four proteins map to the DNA-binding domain in the *Euplotes* telomere protein.

## MATERIALS AND METHODS

### Purification of the *Euplotes* telomere protein

*Euplotes crassus* were grown in septic culture using *Dunaliella salina* as a food source and macronuclei were isolated as previously described (23). The 51 kDa telomere protein was isolated from macronuclei as a DNA-protein complex following exposure to high salt (23). Macronuclei were incubated in 2 M NaCl, 1 mM EDTA, 10 mM Tris (pH 8) for two hours, the resultant nuclear extract was then loaded on a Bio-Gel A15M gel filtration column (Bio-Rad) (23). The DNA-containing fractions were collected, the NaCl removed by dialysis, and the telomere protein released from the macronuclear DNA by extensive digestion with micrococcal nuclease.

### Preparation of the telomere protein for amino acid sequencing

Purified telomere protein was separated from any contaminating histones and micrococcal nuclease by electrophoresis through a 12% SDS polyacrylamide gel. Approximately 1 nmol of protein was electroblotted onto nitrocellulose membrane, the membrane was stained with Ponceau S and the 51 kDa telomere protein band was excised. The excised band was digested extensively with chymotrypsin and the resulting peptide fragments were separated by reverse phase HPLC (30). Seven fragments were then sequenced by Edman degradation.

### Southern blot hybridization

Macronuclear DNA was fractionated in 0.8% agarose gels, transferred to nylon membrane (MCI) using a vacuum blot apparatus (Pharmacia), and fixed to the membrane by baking at 80°C for two hours. For blots probed with the *Oxytricha* $\alpha$ and $\beta$ subunit genes, hybridization was performed overnight at 32°C in 45% formamide, 4× SSC, 5× Denhardt's solution, 0.1% SDS, 200 $\mu$g herring sperm DNA. The blots were washed at 52°C in 4× SSC, 0.1% SDS. For blots probed with oligo #7, hybridization was performed overnight at 47°C in 5× SSC, 5× Denhardt's solution, 0.1% SDS, 200 $\mu$g herring sperm DNA. The blots were washed at 52°C in 5× SSC, 0.1% SDS.

### Library construction and gene cloning

*Euplotes* macronuclear DNA was fractionated in a 0.8% low melting point agarose gel. The DNA in the 1.4−1.6 kb size range was isolated from the gel, G-tailed using terminal transferase, and annealed with C-tailed plasmid vector (pTZ19R, Pharmacia). The resultant mixture was used to transform *E.coli* JM109. The library was screened by colony hybridization (31) using a labeled degenerate oligonucleotide probe that corresponded to a region of the sequenced 51 kDa telomere protein. Positive colonies were further characterized by restriction mapping and Southern blot analysis (31). Double-stranded DNA sequencing was performed by the dideoxy chain termination method using Sequenase (USB).

### Computer analysis

Sequence analyses were performed using the GCG programs (version 7.0) released from the Genetics Computer Group, Madison, Wisconsin. Protein sequence searches were done through the EMBL/Genbank nucleic acid sequence libraries and other protein data banks. Secondary structure prediction and amino acid composition analysis were done using the PROSIS programs (Hitachi America, Ltd.).

## RESULTS

### Identification of genes encoding *Euplotes* telomere-binding proteins

Genes encoding *Euplotes* telomere-binding proteins were identified by screening Southern blots of *Euplotes* macronuclear DNA with the genes encoding the $\alpha$ and $\beta$ subunits of the *Oxytricha* telomere protein. As illustrated by Fig. 1, the gene encoding the $\alpha$ subunit of the *Oxytricha* protein appeared to hybridize to a single band of 1.6 kb while the gene encoding the $\beta$ subunit hybridized to bands of 1.2 and 1.0 kb respectively. To determine which of these DNA molecules encode the *Euplotes* 51 kDa telomere protein, we sequenced peptide fragments from the purified telomere protein and used the sequence of the three longest fragments to design hybridization probes for the corresponding gene. When the degenerate probes were used to screen Southern blots of *Euplotes* macronuclear DNA, the probe corresponding to fragment #7 (oligo #7) hybridized to a single band of 1.6 kb (shown in Fig. 2a lane 2). The probe corresponding to fragment #1 hybridized to the same 1.6 kb band in addition to several others, while the probe corresponding to fragment #5 hybridized to many different bands (data not shown). As the *Oxytricha* $\alpha$ subunit probe and two of the

oligonucleotide probes all hybridized to a DNA molecule of 1.6kb, this molecule was thought to encode the *Euplotes* telomere protein.

Subsequent restriction mapping of the 1.6 kb molecule with Hind III and Xho I revealed that the *Oxytricha* α subunit probe was in fact hybridizing to two separate species of DNA, both of which were 1.6 kb in size. One of the 1.6 kb DNAs was cleaved with Hind III but not Xho I, while the other was cleaved with Xho I but not Hind III (Fig. 2b). The oligo #7 probe hybridized most strongly to the DNA that was cleaved by Xho I (1.6 kb-Xho) (Fig.2a). Thus, this DNA was presumed to encode the 51 kDa telomere protein. Note that the *Oxytricha* α subunit probe and the oligo #7 probe hybridized to opposite ends of the 1.6 kb-Xho molecule. The *Oxytricha* probe recognized a 0.9 kb fragment that was later shown to map to the 5′ end of the gene, while the oligo #7 probe hybridized to a 0.7 kb fragment from the 3′ end of the gene. Although the identity of the DNA cleaved by Hind III (1.6 kb-Hind) was unknown, the apparent similarity in sequence between this DNA molecule and the *Oxytricha* α subunit gene suggested that it might encode a telomere protein homolog.

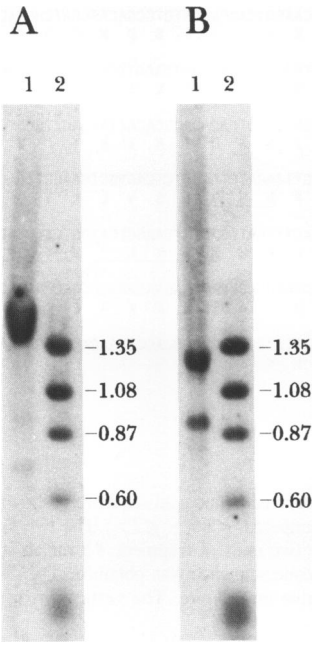


**Figure 1.** Southern blots showing hybridization of the *Oxytricha* α and β subunit genes to *Euplotes* macronuclear DNA. Panel A, *Euplotes* macronuclear DNA probed with the α subunit gene. Panel B, *Euplotes* macronuclear DNA probed with the β subunit gene. Lane 1, *Euplotes* macronuclear DNA; Lane 2, marker DNA. The sizes (in kb) of the markers are indicated at the right.

Clones containing either the 1.6 kb-Xho molecule or the 1.6 kb-Hind molecule were isolated from a *Euplotes* macronuclear DNA library and sequenced. When the cloned 1.6 kb molecules were used to probe Southern blots of restriction digested *Euplotes* macronuclear DNA, under high stringency conditions each molecule hybridized to a different subset of the DNAs identified by the *Oxytricha* α subunit probe. The 1.6 kb-Xho clone hybridized only to the DNA identified by the oligo #7 probe (data not shown) while the 1.6 kb-Hind clone hybridized to the rest of the DNA identified by the *Oxytricha* probe (Fig 2c).

## The gene encoding the 51 kDa telomere protein

The cloned 1.6 kb-Xho DNA molecule had one long open reading frame of 438 amino acids that terminated in UAA, the standard termination codon for *Euplotes* (32−36). This open reading frame appeared to encode the telomere protein because the predicted amino acid sequence matched the partial sequences obtained from the purified telomere protein very closely; 48 out of 50 residues were identical (shown in Fig. 3). The two sites where the sequence of the open reading frame and the peptide fragments did not match probably reflect errors in the protein sequencing as the amino acid sequence was rather indistinct in several places. While ciliates genomes frequently contain several versions of a gene (26,27,37,38), we have been unable to detect any other versions of the telomere protein gene by sequencing PCR amplified RNA.

Although the 438 amino acid open reading frame would encode a protein of 50.5 kd, there was no obvious AUG initiation codon. Upon closer examination of the gene sequence, we detected a

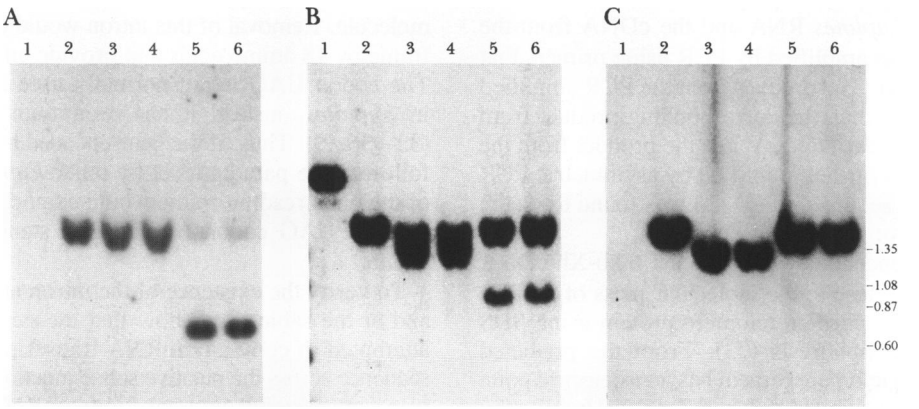


**Figure 2.** Southern blots showing hybridization of the degenerate oligonucleotide probe #7 (panel A), the *Oxytricha* α subunit gene (panel B), and the 1.6 kb-Hind molecule (panel C), to restriction digested *Euplotes* macronuclear DNA. Lane 1, undigested *Oxytricha* macronuclear DNA (15 μg) ; Lane 2, undigested *Euplotes* macronuclear DNA (15 μg); Lane 3, *Euplotes* macronuclear DNA (15 μg) digested with 28 units of Hind III; Lane 4, *Euplotes* macronuclear DNA (15 μg) digested with 56 units of Hind III; Lane 5, *Euplotes* macronuclear DNA (15 μg) digested with 40 units of Xho I; Lane 6, *Euplotes* macronuclear DNA (15 μg) digested with 80 units of Xho I. All restriction digests were performed at 37°C for two hours. The positions and sizes (in kb) of marker DNAs are indicated at the right.

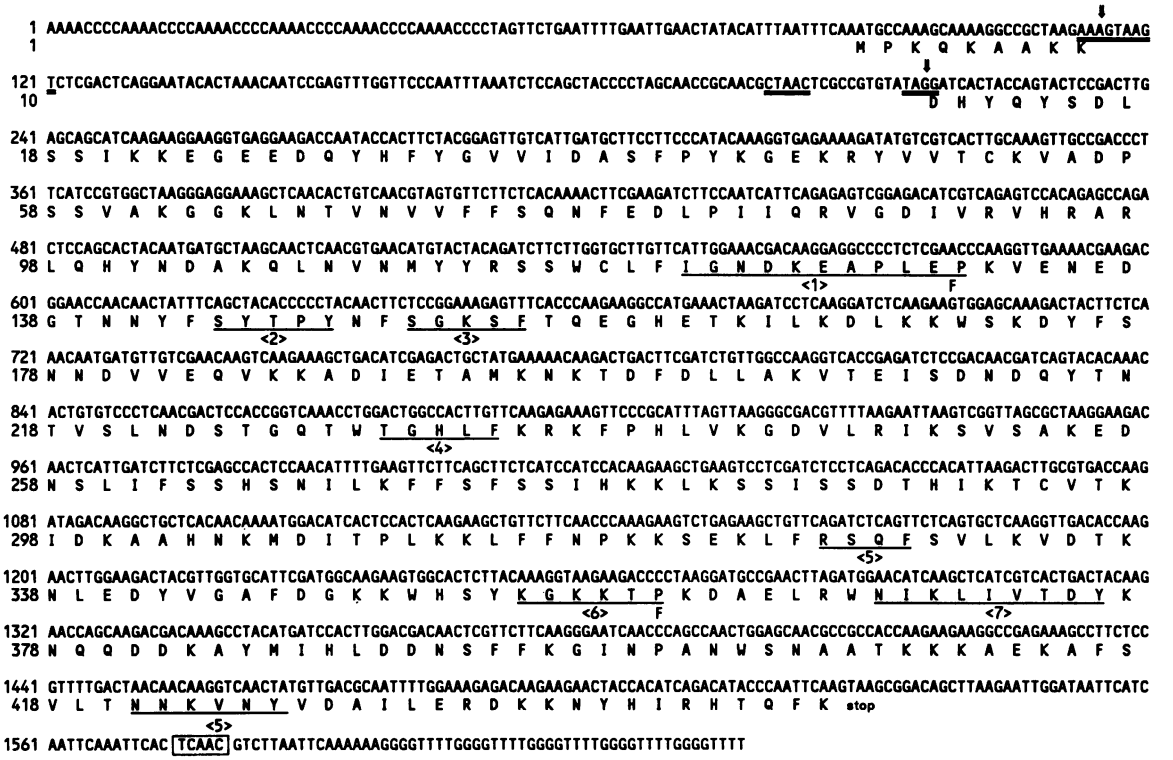


**Figure 3.** Nucleotide sequence of the 1.6 kb-Xho molecule that encodes the *Euplotes* telomere protein. The derived amino acid sequence is shown below the nucleotide sequence. The amino acids corresponding to the sequenced chymotryptic peptides are underlined and numbered <1>–<7>. The two sequence discrepancies are shown as letters below the peptide sequence in fragment #1 (EP→F) and fragment #6 (P→F). The two ends of fragment #5 match separate regions of the protein sequence. This is because the sample applied to the sequencer contained two peptides, so a composite sequence was obtained. The double underlines mark the regions that match the eukaryotic consensus sequence for the 5′ and the 3′ splice sites and the putative branch site. The vertical arrows mark the beginning and the end of the intron. A putative polyadenylation signal is boxed.

putative 101 bp intron near the 5′ end of the gene (marked in Fig. 3). Removal of this intron would generate an open reading frame that started with a methionine and contained 446 amino acids. To verify the existence of the intron, we determined the sequence of the mRNA transcript of the telomere protein gene in the vicinity of the putative splice junction. cDNA was synthesized from total *Euplotes* RNA and the cDNA from the telomere protein gene was amplified by PCR using primers that flanked the putative intron. The product from the PCR amplified RNA was 100 bp smaller than the corresponding product from PCR amplified macronuclear DNA. When the product from the PCR amplified RNA was further amplified by asymmetric PCR and then sequenced, the sequence (Fig. 4a) was found to match the predicted sequence of the ligated exons.

The telomere protein encoded by the cloned 1.6 kb-Xho DNA is 51.5 kDa, this is very close to the molecular mass of 50 kDa that was measured for the purified telomere protein using SDS polyacrylamide gel electrophoresis (23). From the predicted amino acid sequence, the telomere protein has an isoelectric point of 10.2 and is very lysine rich (14% lys, 3% arg). Hydropathy analysis revealed that the N-terminal third of the protein is composed of alternating hydrophobic and hydrophilic sections, while the C-terminal two thirds is predominantly hydrophilic (data not shown)

## The gene encoding the telomere protein homolog

The cloned 1.6 kb-Hind DNA had one long open reading frame of 362 amino acids that terminated in two UAG codons (Fig. 5). This DNA resembled the 1.6 kb-Xho DNA in that there was no obvious AUG initiation codon at the start of the long open reading frame but there appeared to be an intron near the 5′ end of the molecule. Removal of this intron would extend the open reading frame by 18 amino acids and provide an AUG initiation codon. The codon UAG is not normally used as a termination codon in *Euplotes*, instead it has been found to code for cysteine (32–36,39). Thus, if the gene encoded by the 1.6 kb-Hind DNA followed the paradigm set by other *Euplotes* genes, the 3′ end of the open reading frame would extend 88 amino acids beyond the two UAG codons to end at a standard UAA termination codon.

To verify the existence of the intron at the 5′ end of the gene, and at the same time show that the gene was transcribed, we attempted to copy any mRNA transcripts into cDNA and then sequence across the putative splice junction. As before, the cDNA was synthesized from total *Euplotes* RNA (isolated from whole cells) and amplified by asymmetric PCR, the PCR product was then sequenced. The sequence of the PCR product (Fig. 4b) matched the predicted sequence of the ligated exons exactly, demonstrating that the gene was transcribed and the intron

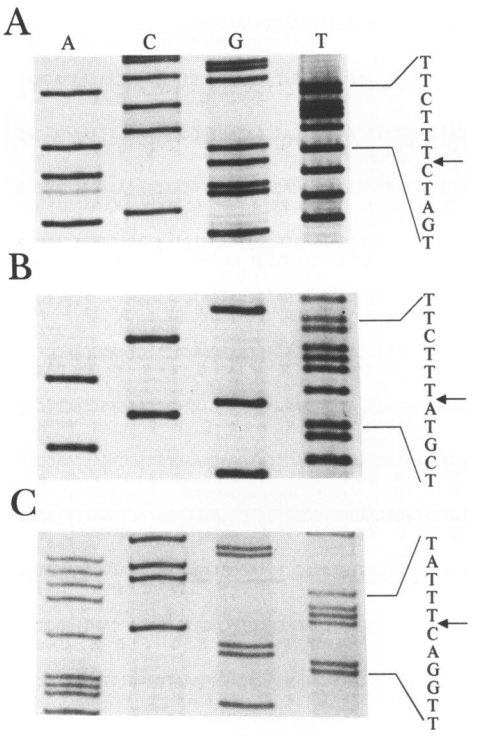


**Figure 4.** Sequence of cDNA in the vicinity of the putative splice sites. Panel A, the 1.6 kb-Xho molecule. Panel B, the 1.6 kb-Hind molecule (the 5′ intron). Panel C, the 1.6 kb-Hind molecule (the 3′ intron). The cDNA sequence shown in all three panels corresponds to the strand that is complimentary to both the mRNA and the DNA sequences shown in Figures 3 and 5. The arrowheads indicate the splice junctions.

removed. We also sequenced the amplified cDNA in the vicinity of the two UAG codons to check for RNA editing or other unusual phenomenon. To our surprise we found that the two UAG codons comprised part of a 24 nucleotide intron. The cDNA sequence (Fig 4c) matched the sequence of the gene 5′ and 3′ of the region containing the UAG codons, but a section of 24 bases that spanned the two UAG codons was completely absent. The 24 nucleotide in-frame intron appeared to be spliced out with high efficiency as the intron was removed in all the RNA preparations examined and the sequence across the splice junction was always extremely clear. Closer examination of the intron sequence revealed both 5′ and 3′ splice sites that closely match the eukaryotic consensus sequences (40); 8 out of 9 nucleotides matched at the 5′ splice site and 5 out of 5 nucleotides at the 3′ splice site (marked in Fig. 5). There is also a UUAAC sequence within the intron that resembles the branch sequences found in higher eukaryotes (41). The distance between the putative branch sequence and the AG at the 3′ end of the intron is 5 nucleotides. This is longer than the minimum distance of 3 nucleotides that has been found in *Schizosaccharomyces pombe* (42).

From the above sequence data it appears that the 1.6 kb-Hind DNA encodes a 460 amino acid protein of 53 kDa. The protein is much less basic than the 51 kDa telomere protein as it has an isoelectric point of 6.2. Hydropathy analysis showed that the protein is generally quite hydrophilic but revealed no other striking features.

## Structure of the genes encoding the telomere protein and telomere protein homolog

The two 1.6 kb molecules have various structural features that are characteristic of genes from hypotrichous ciliates (26,27,32,33, 35−37,39). The 5′ and 3′ noncoding regions are extremely short (35−64 bp), are very AT rich (69−82%) and lack conventional DNA and RNA regulatory signals such as TATA boxes, CAAT boxes, or canonical AATAAA polyadenylation signals (Figs. 3 & 5). The 3′ noncoding regions do contain the sequence TC/TAAC which fits the TYAAC consensus sequence that has been proposed to signal polyadenylation in hypotrichs (37).

The 5′ and 3′ splice sites from all three introns follow the GT-AG rule for intron boundary sequences (40). The 5′ splice sites match the general eukaryotic consensus sequence at all positions except one. However, the single base change of G → A at position −1 generates the consensus sequence for viral 5′ splice sites (43). While the 3′ splice sites match the general eukaryotic consensus sequence exactly, as in many ciliate genes, the polypyrimidine tract preceding the splice site is missing (26,27,35,37,39,44).

The length of telomeric DNA at each end of the 1.6 kb-Xho molecule is different. At the 3′ end there are five repeats of $T_4G_4$ which, given the cloning strategy followed, reflects the standard length of a macronuclear telomere (13). However, at the 5′ end there are 6 repeats of $T_4G_4$. Although telomere length is usually very tightly controlled in *Euplotes*, some variability is occasionally observed (45).

## Sequence conservation between the *Euplotes*, *Oxytricha* and *Stylonychia* telomere proteins and the *Euplotes* telomere protein homolog

When the sequence of the *Euplotes* telomere protein, the α subunit of the *Oxytricha* telomere protein and the *Euplotes* telomere protein homolog were compared, it was immediately apparent that many amino acids were conserved between the three proteins (Fig. 6). The amino acid sequence identity between all three proteins is between 35% and 36% while the level of similarity is between 54% and 56%. For example, the *Oxytricha* and *Euplotes* telomere proteins are 36% identical, the two *Euplotes* proteins are 35% identical, while the *Oxytricha* telomere protein and the *Euplotes* telomere protein homolog are 35% identical. Very similar levels of sequence identity (36−37%) and sequence similarity (56%) were observed when the comparison was extended to the proposed α subunit of the *Stylonychia* telomere protein (Fig. 6). There are multiple regions where at least three or more contiguous amino acids are identical in all the proteins. Many other amino acids are identical in two or three of the proteins while the other protein(s) displays a conservative substitution (eg. ile→leu, lys→arg etc.). A high proportion of the identical amino acids are either aromatic or hydrophobic (26% aromatic, 46% aromatic or hydrophobic).

It is striking that nearly all the extended regions of amino acid sequence identity are within the N-terminal two thirds of the *Euplotes* and *Oxytricha* proteins. It is precisely this portion of the *Euplotes* telomere protein that comprises a distinct structural domain which contains the DNA-binding site (29). Since the *Euplotes* and *Oxytricha* telomere proteins bind similar telomeric DNA sequences, and the resultant DNA-protein complexes yield similar footprints with dimethylsulfate, it is to be expected that

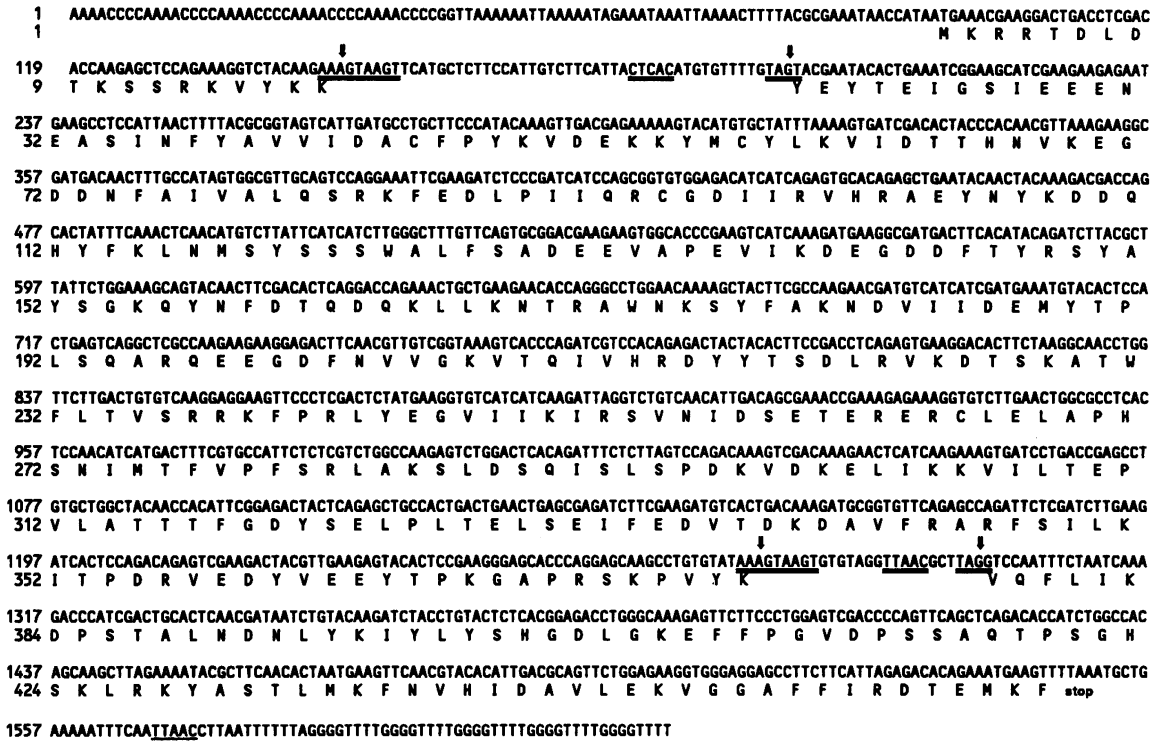

**Figure 5.** Nucleotide sequence of the 1.6 kb-Hind molecule that encodes the *Euplotes* telomere protein homolog. The derived amino acid sequence is shown below the nucleotide sequence. The double underlines mark the regions that match the eukaryotic consensus sequence for the 5′ and the 3′ splice sites and the putative branch sites for each intron. The vertical arrows mark the beginning and the end of the introns. A putative polyadenylation signal is underlined.

the two proteins might have a similar DNA-binding motif (21–23). The extensive sequence conservation between the α subunit of the *Oxytricha* telomere protein and the DNA-binding domain of the *Euplotes* telomere protein lends credence to this idea.

The computer program 'Motifs' was used to search for possible DNA-binding motifs within the *Euplotes* telomere protein and the telomere protein homolog. Although both proteins contain many leucines within the conserved regions, as well as three or four cysteines and multiple histidines, no leucine zipper or zinc finger binding motifs were found. Computer searches of EMBL/Genbank and other protein data banks revealed no obvious sequence similarity between the *Euplotes* telomere protein or the *Euplotes* telomere protein homolog and any proteins other than the α subunits of the *Oxytricha* and *Stylonychia* telomere proteins. Proteins that were scrutinized particularly carefully include those known to bind telomeric DNA such as the β subunit of the *Oxytricha* and *Stylonychia* telomere proteins, yeast RAP1, and various intermediate filament proteins including lamins A, B and C (46).

## DISCUSSION

### The telomere protein sequence

We have shown that *Euplotes crassus* has two 1.6 kb macronuclear DNA molecules that hybridize to the gene encoding the α subunit of the *Oxytricha* telomere protein. Only one of these molecules encodes the 51 kDa telomere protein (the 1.6 kb-Xho molecule) while the other appears to encode a homolog to the telomere protein (the 1.6 kb-Hind molecule). From the sequence of the 1.6 kb-Xho DNA, we have been able to predict the sequence of the entire 51 kDa telomere protein. This amino acid sequence information should be extremely useful for future studies of telomere protein structure and function. Although both the *Oxytricha* and *Euplotes* telomere proteins have some very unusual DNA-binding characteristics, very little is known about how either protein recognizes and binds the terminus of telomeric DNA sequences. Now that the DNA-binding domain of the *Euplotes* protein has been isolated and highly conserved amino acids have been identified within this domain, it will be possible to devise rational experiments that test the contribution of specific amino acids to DNA binding.

It is especially interesting that so many of the amino acids which are conserved between the *Euplotes* and *Oxytricha* telomere proteins are hydrophobic (33%) and/or aromatic (18%). The capacity of both proteins to remain bound to telomeric DNA in the presence of high salt suggests that there is a large non-electrostatic component to the DNA-protein interactions. The composition of the conserved amino acids fits well with this idea. In particular, the large number of aromatic amino acids suggests that one component of the binding reaction could be stacking of phe, trp or tyr with bases in the 3′ single-stranded extension on the telomeric DNA. This type of interaction has been observed with other single-strand DNA-binding proteins such as Gene 5 protein, Gene 32 protein, and *E.coli* single-strand binding protein (47–50). The interactions between the *E.coli* single-strand binding protein and DNA are very salt stable as the protein is not dissociated from poly dT by 5 M NaCl (51).

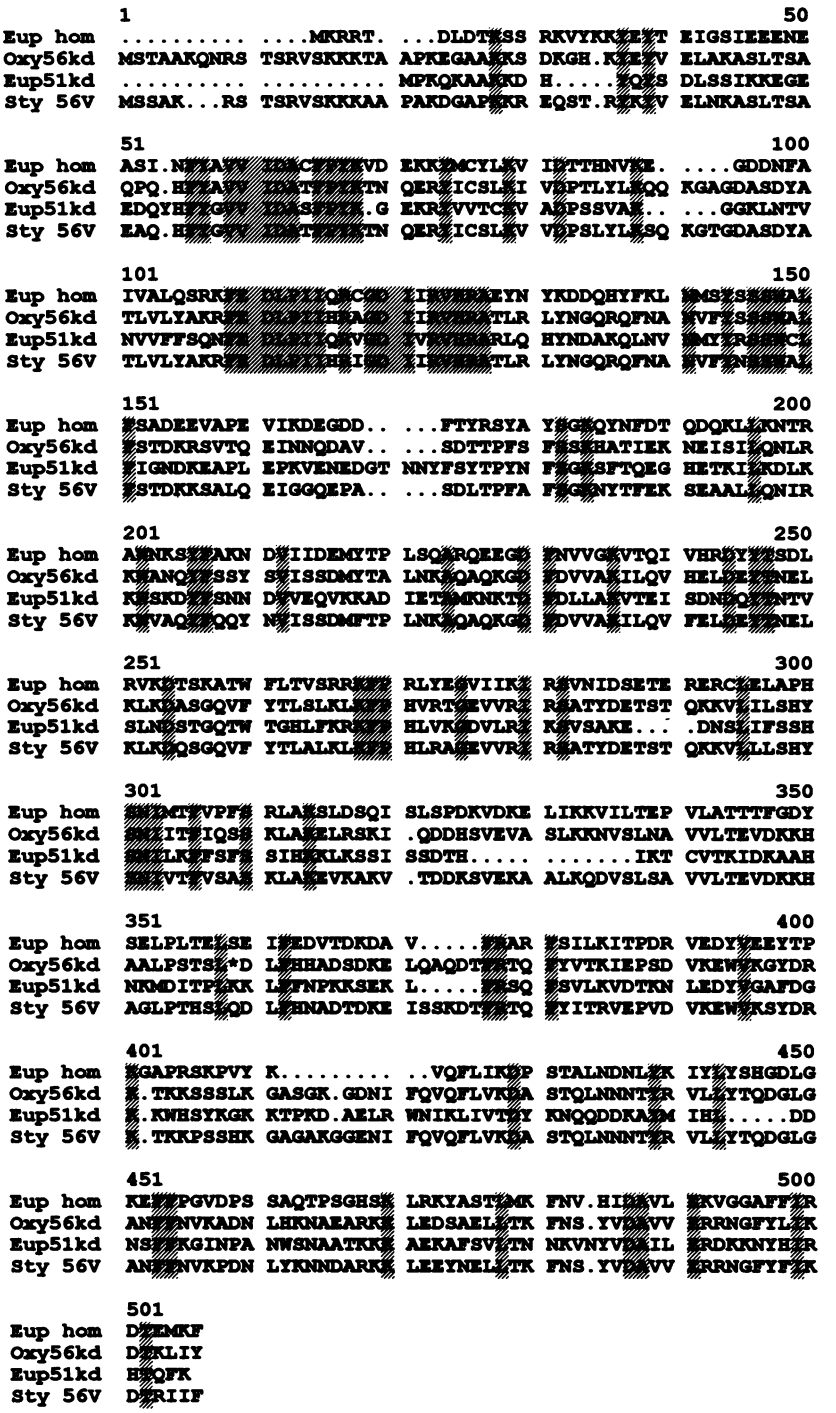

**Figure 6.** Amino acid sequence alignment of the *Euplotes* telomere protein, the α subunit of the *Oxytricha* and *Stylonychia* telomere proteins, and the *Euplotes* telomere protein homolog. Identical residues are highlighted. The gaps marked by dots were introduced to optimize the alignment. The alignment was performed using the GCG program 'Pileup'.

Although certain regions within the N-terminus of the *Euplotes* and *Oxytricha* telomere proteins are very highly conserved, the overall level of sequence conservation (36% identical amino acids) is not particularly high compared to other structural proteins from the same ciliates. For example, the amino acid sequence of actin from *Euplotes* and *Oxytricha* is 61% identical while those regions of histone H4 that have been sequences are 95% identical (38). The significant sequence divergence between the two telomere proteins may explain their structural differences. For example, the tight association of the α and β subunits in the *Oxytricha* protein as compared to the apparent lack of association of the *Euplotes* protein with a β subunit (23,24,29).

Fairly extensive efforts to isolate additional subunits of the *Euplotes* telomere protein have proved unsuccessful (C.Price

unpublished results) indicating that a β subunit is unlikely to be permanently associated with the 51 kDa α subunit. However, we have shown that the *Euplotes* macronucleus contains genes with homology to the *Oxytricha* β subunit (see Figure 1b). Thus, *Euplotes* may well have β subunit homologs, but they are likely to differ significantly from the β subunit of the *Oxytricha* protein.

### The telomere protein homolog

The discovery that *Euplotes* has two genes which are similar to the *Oxytricha* α subunit gene was quite unexpected as only one gene had previously been identified in either *Oxytricha* or *Stylonychia*. The second *Euplotes* gene (1.6 kb-Hind) does not appear to be a pseudogene as it is transcribed and the resultant transcript has a long open reading frame with normal translation start and stop signals. Detection of the 1.6 kb-Hind DNA required careful manipulation of the hybridization stringency during Southern blot analysis. Thus, it is possible that equivalent genes exist in *Oxytricha* and *Stylonychia* but they have not yet been detected.

The 1.6 kb-Hind DNA clearly encodes a homolog of the telomere protein as there is extensive sequence homology between the proposed protein sequence and the sequence of the *bona fide* telomere proteins from both *Oxytricha* and *Euplotes*. It is particularly striking that so many of the conserved amino acids are identical in all three proteins and that all the extended regions of conservation map within the DNA-binding domain of the *Euplotes* telomere protein (29). These observations strongly suggest that the telomere protein homolog may also bind telomeric $T_4G_4$ sequences.

At present it is not clear why *Euplotes* should have several different proteins that bind telomeric DNA. A possible reason is that one protein binds macronuclear telomeres while the other binds either micronuclear telomeres or the telomeres from developing macronuclei. The telomeres from micronuclei and developing macronuclei have longer stretches of $C_4A_4.T_4G_4$ repeats than telomeres from mature macronuclei. If the length of the telomeric DNA is recognized by (or even determined by) telomere-binding proteins, it may be necessary to have a set of proteins that all recognize the $C_4A_4.T_4G_4$ sequence via a conserved DNA-binding site, but which differ in the regions that detect DNA length.

### The 24 nucleotide intron

The size range of pre-mRNA introns varies from species to species and it is well established that certain organisms have an abundance of very small introns. For example, >60% of introns from *Schizosaccharomyces pombe* are less than 100 nucleotides in length, about 50% of the introns from *Tetrahymena thermophila*, and 62% (five out of eight) in *Euplotes* species (35,39,42,44). However, for most organisms the lower limit on intron size appears to be in the 35–50 nucleotide range, although introns as short as 31 nucleotides have been found in *Drosophila* (42,43,52). Given the large number of snRNPs that are thought to pair simultaneously with an intron and with each other, it has generally been assumed that a length of 31 nucleotides is the minimum that can include all the sequences necessary for spliceosome assembly. Thus, the discovery of an efficiently spliced intron that is only 24 nucleotides in length is extremely surprising. It will be interesting to determine whether all of the snRNPs and other proteins that are thought to interact directly with larger introns are needed to splice the 24 nucleotide intron.

## ACKNOWLEDGEMENTS

We thank Tom Cech for providing clones of the genes encoding the α and β subunits of the *Oxytricha* telomere protein, Angela Adams for supplying *Euplotes* RNA, and DeWight Williams for growing *Euplotes crassus* and isolating macronuclei. We thank Judith Berman and Dorothy Shippen-Lentz for reading this manuscript and making helpful comments and we thank Scott Perez for his assistance with the figures. The research was supported by the National Institute of Health (grant # GM41803), a Basil O'Connor Starter Scholar Research Award from the March of Dimes Birth Defects Foundation (CMP) and an American Cancer Junior Faculty Research Award (CMP).

## REFERENCES

1. Zakian, V. A. (1989) *Ann. Rev. Genet.*, **23**, 579–604.
2. Blackburn, E. (1991) *Nature*, **350**, 569–573.
3. Biessmann, H. and Mason, J. M. (1992) submitted to *Adv. Genet.*
4. Yu, G-L., Bradley, J. D., Attardi, L. D. and Blackburn, E. H. (1990) *Nature*, **344** 126–132.
5. de Lange, T. (1992) *EMBO J.*, **11**, 717–724.
6. Chiou, S. and Blackburn E. H. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 2263–2267.
7. Edwards, C. A. and Firtel, R. A. (1984) *J. Mol. Bio.*, **180**, 73–90.
8. Gottschling, D. E. and Cech, T. R. (1984) *Cell*, **38**, 501–510.
9. Budarf, M. L. and Blackburn E. H. (1986) *J. Biol. Chem.*, **261**, 363–369.
10. Lucchini, R. U., Pauli, U., Braun, R., Koller, T. and Sogo, J. M. (1987) *J. Mol. Biol.*, **196**, 829–843.
11. Wright, J. H., Gottschling, D. E. and Zakian, V. A. (1992) *Genes Dev.*, **6**, 197–210.
12. Price, C. M. (1992) *Current Opinion in Cell Biol.*, **4**, 379–384.
13. Klobutcher, L. A., Swanton, M. T., Donini, P. and Prescott, D. M. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 3015–3019.
14. Henderson, E. R. and Blackburn, E. H. (1989) *Mol. Cell. Biol.*, **9**, 345–348.
15. Coren, J., Epstein, E. and Vogt, V. (1991) *Mol. Cell Biol.*, **11**, 2282–2290.
16. Liu, Z. and Tye, B. (1991) *Genes Dev.*, **5**, 49–59.
17. Lustig, A. J., Kurtz, S. and Shore, D. (1990) *Science*, **250**, 549–553.
18. Conrad, M. N., Wright, J. H., Wolf, A. J. and Zakian, V. A. (1990) *Cell*, **63**, 739–750.
19. Sussel, L. and Shore, D. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 7749–7753.
20. Hardy, C. F. J., Sussel, L. and Shore, D. (1992) *Genes Dev.*, **6**, 801–814.
21. Gottschling, D. E. and Zakian, V. A. (1986) *Cell*, **47**, 195–205.
22. Price, C. M. and Cech, T. R. (1987) *Genes Dev.*, **1**, 783–793.
23. Price, C. M. (1990) *Mol. Cell Biol.*, **10**, 3421–3431.
24. Price, C. M. and Cech, T. R. (1989) *Biochem.*, **28**, 769–774.
25. Klobutcher L. A. and Prescott. D. M. (1986) In Gall, J. G. (ed.). Molecular biology of ciliated protozoa. Academic Press, New York. pp.111–154.
26. Gray, J. T., Celander, D. W., Price, C. M. and Cech, T. R. (1991) *Cell*, **67**, 807–814.
27. Hicke, B., Celander, D., Macdonald, G., Price, C. and Cech, T. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 1481–1485.
28. Fang, G. and Cech, T. R. (1991) *Nucleic Acids Res.*, **19**, 5515–5518.
29. Price, C. M., Skopp, R. Krueger, J. and Williams, D. (1992) *Biochemistry.*, In press.
30. Matsudira, P. T. (1989) A practical guide to protein and peptide purification for microsequencing. Academic Press.
31. Sambrook. J., Fritsch, E. F. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Lab., Cold Spring Harbor, NY.
32. Harper, D. S. and Jahn, C. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 3252–3256.
33. Miceli, C., La Terza, A. and Melli, M. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 3016–3020.
34. Meyer, F., Schmidt, H., Plumper, E., Hasilik, A., Mersmann, G., Meyer, H., Engstrom, A. and Heckmann, K. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 3758–3761.
35. Meyer, F., Schmidt, H. J. and Heckman, K. (1992) *Developmental Genetics*, **13**, 16–25.
36. Hauser, L. J., Roberson, A. E. and Olins, D. E. (1991) *Chromosoma*, **100**, 386–394.
37. Williams, K.R. and Herrick, G. (1991) *Nucleic Acids Res.*, **17**, 4717–4724.

38. Harper, D. S. and Jahn, C. L. (1989) *Gene*, **75**, 93–107.
39. Brunen-Nieweler, C., Schmidt, H. J. and Heckmann, K. (1991) *Gene*, **109**, 233–237.
40. Mount, S. M. (1982) *Nucleic Acids Res.*, **10**, 459–472.
41. Green, M. R. (1986) *Ann. Rev. Genet.*, **20**, 671–708.
42. Prabhals, G., Rosenberg, G. H. and Kaufer, N. F. (1992) *Yeast*, **8**, 171–182
43. Shapiro, M. B. and Senapathy, P. (1987) *Nucleic Acids Res.*, **15**, 7155–717
44. Csank, C., Taylor, F. M. and Martindale, D. W. (1990) *Nucleic Acids Res.*, **18**, 5133–5141.4.
45. Klobutcher, L. A., Turner, L. R. and Peralta, M. E. (1991) *J. Protozool*, **38**, 425–427.
46. Shoeman, R. L., Wadle, D., Scherbarth, A. and Traub, P. (1988) *J. Biol. Chem.*, **263**, 18744–18749.
47. Chase, J. W. and Williams, K. R. (1986) *Ann. Rev. Biochem.*, **55**, 103–136.
48. Dick, L. R., Geraldes, C. F., Sherry, A. D., Gray, C. W. and Grey, D. M. (1989) *Biochem.*, **28**, 7896–7904.
49. King, G. C. and Coleman, J. E. (1987) *Biochem.*, **26**, 2929–2937.
50. King, G. C. and Coleman, J. E. (1988) *Biochem.*, **27**, 6947–6953.
51. Lohman, T. M. and Overman, L. B. (1985) *J. Biol. Chem.*, **260**, 3594–3603.
52. Hawkins, J. D. (1988) *Nucleic Acids Res.*, **16**, 9893–9908.