Nucleic Acids Research, 1998, Vol. 26, No. 7 **1731–1740**

Sequence-specific DNA recognition by the Myb-like domain of the human telomere binding protein TRF1: a model for the protein–DNA complex

Peter König[+], Louise Fairall and Daniela Rhodes*

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Received November 28, 1997; Revised and Accepted February 9, 1998

ABSTRACT

Telomeres consist of tandem arrays of short G-rich sequence motifs packaged by specific DNA binding proteins. In humans the double-stranded telomeric TTAGGG repeats are specifically bound by TRF1 and TRF2. Although telomere binding proteins from evolutionarily distant species are not sequence homologues, they share a Myb-like DNA binding motif. Here we have used gel retardation, primer extension and DNase I footprinting analyses to define the binding site of the isolated Myb-like domain of TRF1 and present a three-dimensional model for its interaction with human telomeric DNA. Our results suggest that the Myb-like domain of TRF1 recognizes a binding site centred on the sequence GGGTTA and that its DNA binding mode is similar to that of the homeodomain-like motifs of the yeast telomere binding protein RAP1. The implications of these findings for recognition of telomeric DNA in general are discussed.

INTRODUCTION

Telomeres are specialized nucleoprotein complexes that protect the ends of linear eukaryotic chromosomes from degradation and end-to-end fusion (1). Loss of a telomere in yeast leads to cell cycle arrest (2) and telomere shortening in human cells has been implicated in both senescence and tumourigenesis (3). In most organisms telomeric DNA consists of G-rich sequence repeats of 6–8 bp arranged in tandem to form long double-stranded arrays (4). These sequence repeats are bound by sequence-specific DNA binding proteins such as RAP1 from the budding yeasts *Saccharomyces cerevisiae* (5) and *Kluyveromyces lactis* (6), the more distantly related TAZ1 from *Schizosaccharomyces pombe* (7) and the vertebrate TTAGGG repeat binding factors TRF1 and TRF2 (8–11). For several of these telomere binding proteins there is *in vivo* evidence that they are involved in telomere length regulation, presumably by regulating access of telomerase, the reverse transcriptase-like enzyme that adds telomeric repeats (7,12–14).

Despite the apparently conserved function of telomeric repeat binding proteins, they share a limited amino acid sequence similarity. This similarity is confined to a domain of ~50 amino acids which resembles the DNA binding motif present in the c-Myb family of transcriptional activators (4). The Myb-like domains of human TRF1 and TRF2 are very closely related to each other (56% sequence identity), as are those of *S.pombe* TAZ1 and human TRF1 (30% sequence identity). Interestingly, whereas these telomere repeat binding proteins contain only a single Myb-like domain, the DNA binding domain of the Myb proteins typically consists of three tandem repeats of the Myb motif, and at least two are required for sequence-specific DNA recognition (15,16). This raises questions concerning the oligomerization state of telomere repeat binding proteins. In the case of human TRF1 it has recently been demonstrated that the protein binds as a pre-formed homodimer to multiple copies of the TTAGGG telomeric repeat (17).

The apparent requirement of two juxtaposed Myb-domains for sequence-specific recognition is observed in the high resolution three-dimensional structure of the DNA binding domain of the telomere binding protein RAP1 from *S.cerevisiae* in complex with a telomeric DNA binding site (18). RAP1 binds to telomeric DNA as a monomer but contains two intramolecular and structurally very similar subdomains which are arranged on the DNA in a tandem orientation, each recognizing the sequence GGTGT, which occurs twice in the DNA-binding site. The structure shows that, despite any significant sequence similarity, the RAP1 subdomains are structurally closely related to the Myb DNA binding motifs (19). This structural similarity arises from the presence in both proteins of a three-helix bundle (4,18). Embedded in each three-helix bundle is a helix–turn–helix (HTH) DNA binding motif (20). The HTH motif sits in the DNA major grove and makes base-specific contacts primarily using residues from the DNA recognition helix. However, the structures of the c-Myb and RAP1 domains are different in one important aspect. In RAP1 the three-helix bundle of each domain is augmented by a short N-terminal arm that interacts directly with bases in the minor groove. This DNA binding mode classifies the RAP1 domains as homeodomains (21), another class of three-helix bundle containing a DNA binding motif. The presence of the N-terminal arm has the effect of increasing the length of DNA sequence that can be recognized from 3–4 bp, as seen the structure of the c-Myb–DNA complex, to 5–6 bp, as seen for RAP1 (18,19).

Since the DNA binding affinity of the TRF1 homodimer increases with the number of TTAGGG telomeric repeats (17,22), binding of the full-length protein is likely to be complex. In order to simplify the problem we have analysed the binding properties of the isolated Myb-like domain of human TRF1 to

*To whom correspondence should be addressed. Tel: +44 1223 402441; Fax: +44 1223 213556; Email: rhodes@mrc-lmb.cam.ac.uk

[+]Present address: Department of Biochemistry and Biophysics, School of Medicine, University of California San Francisco, San Francisco, CA 94143-0448, USA

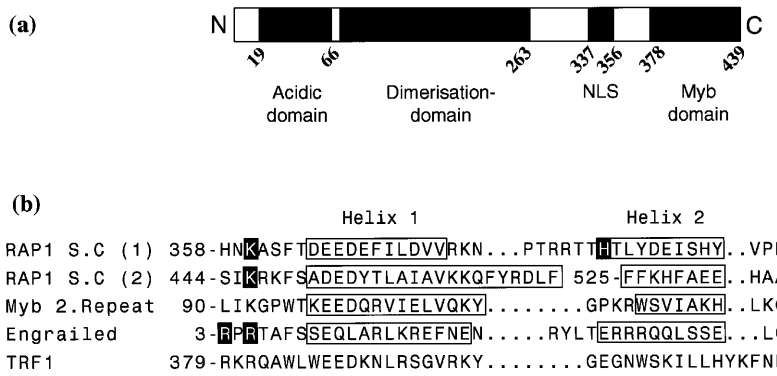

Figure 1. (**a**) The domain structure of the human telomeric repeat binding factor TRF1 (8). The various domains, including that of the Myb-like motif, are indicated. NLS corresponds to the nuclear localization signal. (**b**) Sequence alignment of the Myb-like domain of TRF1 with the second repeat of c-Myb, the two DNA binding domains of RAP1 and the homeodomain of Engrailed. The α-helical regions are boxed and the amino acid side chains involved in contacting DNA bases directly are highlighted.

obtain precise information on the phase and sequence of the DNA binding site recognized by this domain. In particular, we wanted to address the question of whether the TRF1 Myb-like domain binds like c-Myb (4) or as a homeodomain, as seen in RAP1 (21). We show that in contrast to c-Myb, the isolated Myb-like domain of TRF1 binds specifically and with a significant affinity to telomeric DNA as a monomer. Primer extension analyses and DNase I footprinting studies define the binding site of the TRF1 Myb-like domain to be centred on the sequence 5′-GGGTTA-3′ and show that two domains can bind on telomeric DNA with a centre-to-centre spacing of 6 bp. Based on a combined comparison of amino acid sequences and structural information from the structurally closely related RAP1, the Engrailed homeodomain and c-Myb protein–DNA complexes, we propose a three-dimensional model for interaction of the Myb-like domain of TRF1 with DNA. The implications of this model for binding of the full-length TRF1 dimer and emergence of a conserved protein fold for recognition of conserved telomeric repeats are discussed.

MATERIALS AND METHODS

Oligonucleotides

All oligonucleotides were synthesized on an Applied Biosystems 380B DNA synthesizer and purified on denaturing polyacrylamide gels. T3692 and T5630 are complementary oligonucleotides containing the two-repeat telomeric sequence (T3692, 5′-CAGATC-GAATTCGGATTCCCTAACCCTAAGGAAGGAATTCATCC-AG-3′). The complementary oligonucleotides T3693 and T5172 are identical to T3692/T5630 but contain a mutated two-repeat telomeric sequence (5′-…TGCCTAAGCCTAAG…-3′). T5959 and T5960 are complementary oligonucleotides containing the three-repeat telomeric sequence (T5959, 5′-CAGATCGAATTCG-GCCCTAACCCTAACCCTAAAGGAATTCATCCAG-3′). The oligonucleotides used as primers for the primer extension analysis were T3629 (5′-CAGATCGAATTCG-3′) and T6217 (5′-CTGGATGAATTCC-3′). The concentrations of oligo-nucleotides were estimated from OD_{260} measurements using sequence-specific extinction coefficients (32).

Cloning, expression and purification of $TRF1_{371–439}$ and $TRF1_{337–439}$

To obtain the Myb-like domain of TRF1, residues 371–439 and 337–439 of TRF1 (Fig. 1a) were PCR amplified (Vent DNA polymerase; Biolabs) from a plasmid containing the TRF1 gene (a gift from Dr Titia de Lange). Primers compatible for cloning into the *Nde*I and *Bam*HI restriction sites of expression plasmid pET13A (33) were used and the resulting constructs confirmed by dideoxy sequencing. *Escherichia coli* strain BL21 pLysS was transformed (34) and the cells grown at 37°C in 2× TY medium containing 50 µg/ml kanamycin, 25 µg/ml chloramphenicol and 0.4% glucose to an OD_{600} of ~0.3–0.6 before induction with 1 mM IPTG. Cells were then grown for a further 4 h, harvested by centrifugation and the pellets stored at –70°C. The pellets were resuspended in a buffer containing 50 mM Tris–HCl, pH 8.0, 5% glycerol, 5 mM EDTA, 1 mM benzamidine, 1 mM phenylmethyl-sulfonyl fluoride, sonicated and centrifuged. Both $TRF1_{371–439}$ and $TRF1_{337–439}$ proteins were purified using column chromatography. Supernatants were loaded onto a DEAE CL-6B Sepharose column (12 × 5 cm, flow rate 120 ml/h, 4°C; Pharmacia) and the flow-through collected. This was loaded directly onto a HPLC SP-5PW column (21.5 × 1.5 cm, flow rate 4 ml/min, 22°C; Toyosoda) and eluted with a gradient of 0.0–1.0 M NaCl in 50 mM Tris–HCl, pH 8.0. At this stage $TRF1_{371–439}$ was purified further on a Phenyl Sepharose column (8 × 2.5 cm, flow rate 1 ml/min, 22°C; Pharmacia). Prior to loading on the column the sample was made 1 M $(NH_4)_2SO_4$, 0.6 M NaCl. $TRF1_{371–439}$ was eluted with a gradient of 1.0–0.0 M $(NH_4)_2SO_4$ in 50 mM Tris–HCl, pH 8.0, 0.6 M NaCl. The final purification step for both $TRF1_{371–439}$ and $TRF1_{337–439}$ was on a HPLC RP-Nucleosil 300-10 C8 column (15 × 2 cm, flow rate 5 ml/min, 22°C; Macherey-Nagel). Protein samples were made 25% acetonitrile and 1% trifluoroacetic acid before loading onto the column and elution carried out with a gradient of 25–60% acetonitrile in 0.1% trifluoroacetic acid. Both proteins were refolded on a HPLC SP-5PW column (7.5 × 0.75 cm, flow rate 1 ml/min, 22°C; Toyosoda). The buffer was changed to 0.1 M (NH_4)OAc, pH 4.8, 50 mM Tris–HCl, pH 8.0, over 10 min. Proteins were eluted with 1.0 M (NH_4)OAc, pH 4.8, 50 mM Tris–HCl, pH 8.0, and concentrated by ultrafiltration in a Centricon SR3 (Amicon). Aliquots were flash frozen in liquid nitrogen. At the end of the purification proteins were >90% pure as judged by SDS–PAGE and Coomassie staining. The identities

of $TRF1_{371–439}$ and $TRF1_{337–439}$ were confirmed by N-terminal amino acid sequencing. Protein concentrations were estimated from the extinction coefficient (ε_{280} ~ 40727/M/cm) determined from amino acid analysis.

Gel retardation analysis and estimation of dissociation constants

For binding experiments one of the DNA strands was 5′-labelled with [γ-^{32}P]ATP using T4 polynucleotide kinase (Biolabs) and annealed with an equimolar amount of unlabelled complementary strand. Binding reactions were performed at somewhat different protein and DNA concentrations for different experiments. Details are given in the figure legends. The binding buffer contained 20 mM HEPES, pH 7.9, 150 mM KCl, 1 mM $MgCl_2$, 0.1 mM EDTA, 4% Ficoll 400, 1 mM DTT, 5% glycerol and 0.5 mg/ml BSA. Binding reactions were incubated at 4°C for at least 1 h before analysis on 8–10% native polyacrylamide gels (37:1, 0.5× TBE). Electrophoresis was carried out at 4°C. Gels were dried and visualized either by autoradiography or by using a Molecular Dynamics phosphorimager system and the supplied software (ImageQuant). Band intensities were corrected for background and compared with a calibration curve with known DNA concentrations.

Dissociation constants (K_d) were calculated based on the equations of Riggs (35). The dissociation constant for binding of $TRF1_{371–439}$ to the two-repeat DNA site is $K_d = [P][D]/[PD]$, where [PD] is the concentration of protein–DNA complex, [P] that of unbound protein and [D] that of unbound DNA. With $[P] = [aP_{total}] - [PD]$, where $[aP_{total}]$ is the total concentration of protein determined by amino acid analysis multiplied by the activity coefficient a, $[PD_m] = [aP_{total}] - K_d([PD_m]/[D_m]$, where $[PD_m]$ and $[D_m]$ are the measured band intensities compared with a calibration curve with known DNA concentrations.

The ratio of non-specific (u) to specific (s) dissociation constants is $K_u/K_s = [D_u][PD_s]/[D_s][PD_u]$. If the concentration of active protein is significantly lower than that of DNA the terms for unbound DNA, $[D_s]$ and $[D_u]$, can be replaced by their total concentrations, $[D_{s/total}]$ and $[D_{u/total}]$. The concentration of non-specific complex $[PD_u]$ corresponds to the measurable difference between specific complex without competition $[PD_{s/m}]°$ and with competition by non-specific DNA $[PD_{s/m}]$, resulting in $D_{u/total_1}]/[D_{s/total_1}] = K_u/K_s\ ([PD_{s/m}]° - [PD_{s/m}])/[PD_{s/m}]$.

Primer extension analysis

To generate DNA fragments that are truncated at different positions from both ends of the double-stranded DNA fragment 10 pmol primers T3629 and T6217 were radioactively labelled at their 5′-ends, annealed with 15 pmol templates T3692 or 5959 and T3629 or T5630 (see above) respectively, ethanol precipitated, dried and resuspended in TE (10 mM Tris–HCl, pH 8.0, 1 mM EDTA). Elongation of the primers was carried out using the didioxy sequencing reaction. DNA (160 µl) was mixed with 68 µl water, 48 µl 100 mM DTT, 32 µl buffer (280 mM Tris–HCl, pH 7.5, 350 mM NaCl, 100 mM $MgCl_2$), 16 µl Sequenase buffer and 4 µl 13 U/ml Sequenase (US Biochemical–Amersham). The mixture was divided into four 80 µl samples and each incubated with 40 µl T, C, G, A termination mixes (0.015 mM dXTP, 0.015 mM ddXTP, 0.075 mM dNTP, 40 mM Tris–HCl, pH 7.5, 10 mM $MgCl_2$, 50 mM NaCl) for 5 min at 37°C and phenol/chloroform extracted. Termination mixes were combined and the overhanging 5′-ends of the templates removed by addition of 1.9 ml 124 U/µl mung bean nuclease (Pharmacia). DNA samples were then phenol/chloroform extracted, concentrated by rotary evaporation to a volume of 100 µl and dialysed in TE. Samples of 12 µl DNA were incubated with protein at a concentration of 1.3×10^{-8}–3.2×10^{-8} M in 30 µl and binding reactions carried out as described above. Bound DNA was separated from unbound on native polyacrylamide gels (10%, 0.5× TBE). Gels were autoradiographed wet, bands corresponding to bound and unbound DNA excised and the DNA extracted in 0.5 M $(NH_4)OAc$, pH 4.8, 0.1% SDS, 1 mM EDTA, 50 µg/ml proteinase K and 10% methanol for 36 h at 37°C. After filtration (Spin-X) and ethanol precipitation pellets were washed and dried. Pellets were resuspended and analysis carried out by electrophoresis in a 15% sequencing polyacrylamide gel. Relative dissociation constants were calculated from the intensity of the bands quantified using a Molecular Dynamics phosphorimager system and the supplied software (ImageQuant). The relative dissociation constant (K_{rel}) for binding to the two-repeat DNA site with length x in comparison with the full-length fragment (36) is $K_{rel} = K_x/K_{46} = [D_x]/[PD_x][PD_{46}]/[D_{46}]$. The relative concentrations correspond to the ratios of the band intensities in the unbound and bound DNA samples. The dissociation constant for the binding of two proteins to the three-repeat DNA site is $K_1K_2 = [D][P]^2/[P_2D]$; the relative dissociation constant $K_{rel} = (K_1K_2)_x/(K_1K_2)_{46} = [D_x]/[[P_2D_x][P_2D_{46}]/[D_{46}]$. In order to have a reliable estimate for $[PD_{46}]/[D_{46}]$ and $[P_2D_{46}]/[D_{46}]$, this ratio was replaced by that of the sums of several DNA fragments with full binding activity.

DNase I footprinting analysis

For the DNase I footprint analysis the two-repeat sequence (annealed oligonucleotides T3692 and T5630) were cloned into the *Sma*I site of pBend2 (36). The *Nhe*I–*Hin*dIII restriction fragment was then cut out and gel purified. The 3′-end of the G-rich strand was radioactively labelled by a filling in reaction of the *Nhe*I site, using [α-^{32}P]dCTP and reverse transcriptase. The 3′-end of the C-rich strand was labelled in the same way by filling in the *Hin*dIII site, using [α-^{32}P]dATP. Binding reactions were carried out at a DNA concentration of ~1×10^{-9} M and a $TRF1_{371–439}$ concentration of 4×10^{-8} M, in a binding buffer containing 20 mM HEPES, pH 8.0, 100 mM KCl, 10 µg/ml BSA, 0.1% Triton X-100, 5% glycerol, 2 mM $MgCl_2$, 10 µg/ml sonicated calf thymus DNA. The naked DNA and protein–DNA complex were digested at room temperature in 20 µl using 2 µl DNase I at 0.5 U/ml. Aliquots of 10 µl were removed at 4 and 8 min and DNase I digestion stopped by addition to Eppendorf tubes containing 10 µl 6 mM EDTA and 20 µl phenol. Samples were then extracted, 1 µg sonicated calf thymus DNA added, made 0.3 M NaOAc, pH 6.0, and ethanol precipitated. Pellets were carefully washed with cold 70% ethanol (–20°C) and dried. Analysis was carried out in 8% polyacrylamide sequencing gels. The gels were dried and visualized using a Molecular Dynamics phosphorimager. The intensities of bands were quantified using the program GELTRAK (37). Difference probability plots were calculated by subtracting the probabilities of cleavage for each bond in the naked DNA from the corresponding probabilities in the protein–DNA complex. The probabilities of cleavage were calculated using the equation:

$$P_n = A_n \Big/ \sum_{m=n}^{m=n_{max}} A_m$$

where the integrated area (A_n) of a band *n* is divided by the sum of the integrated area (A_m) of uncut DNA plus all the fragments (bands) longer than and including band *n* (38). The difference probability plots were calculated by subtracting the natural logarithm of the probability of cutting for the naked DNA from the natural logarithm of the probability of cutting for the protein–DNA complex. In each case the bond cut is on the 3′-side of the base numbered.

Modelling of the of TRF1 Myb-like motif–DNA complex

All model building was done using the program O (39) running on an Indigo 2 workstation (Silicon Graphics). The coordinates of c-Myb (PDB accession no. 1mse), Engrailed (PDB accession no. 1hdd) and RAP1 (PDB accession no. 1ign) were obtained from the Brookhaven Protein DataBank. The side chains of c-Myb (residues c93–c135) were changed to correspond to the TRF1 sequence. This caused only one major steric clash, which was corrected. Two additional residues, Y410 and K411 of TRF1, were inserted in the loop between helices 2 and 3, assuming the most likely backbone conformation. The N-terminal arm of the Engrailed homeodomain (residues 3–5, numbering according to PDB accession no. 1hdd) was appended to the N-terminus of the TRF1 Myb domain model and modified in a single position (Pro4) for the TRF1 sequence. A B-DNA model of the human telomeric sequence was generated using Insight II (Biosym Technologies).

Prediction of the theoretical DNase I footprint

In order to obtain a theoretical DNase I footprint the B-DNA of the modelled protein–DNA complex was extended beyond the region of the TRF1 Myb-like domain–DNA complex by ~10 bp. Then the three-dimensional structure of the DNase I–DNA complex (PDB accession no. 1dnk; 30) was superimposed along the DNA of the model. The superimpositions were done using the C1′ atoms of three base pairs surrounding the proposed active cleavage site of DNase I and the corresponding atoms of the three consecutive base pairs in the modelled complex. Protection from DNase I cleavage was assumed if the physical shape of DNase I sterically clashed with that of the TRF1 Myb-like domain.

RESULTS

The Myb-like domain of TRF1 binds specifically to telomeric DNA as a monomer

The first experiment performed was to determine whether the Myb-like domain of TRF1, which is located in the C-terminus of the 439 residue protein (Fig. 1a; 8), is sufficient to bind to DNA and whether it binds as a monomer or dimer. In order to accomplish this two different length constructs encompassing this domain, $TRF1_{371–439}$ and the longer $TRF1_{337–439}$, were expressed in *E.coli* and the two proteins purified to homogeneity (Materials and Methods). The two different length proteins were incubated at varying concentrations with a DNA binding site containing two copies of the human telomeric TTAGGG repeat and complex formation analysed by native gel electrophoresis. Figure 2 (lanes 1 and 5) shows that each protein binds to the two-repeat sequence and forms a single complex whose migration is dependent on the size of the protein used. Since no intermediary migrating protein–DNA complex was detected when equimolar amounts of the two proteins

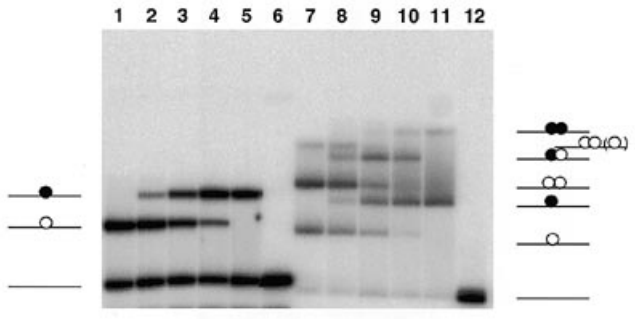

Figure 2. Gel retardation analysis of the Myb-like domain of TRF1 to $(TTAGGG)_2$ and to $(TTAGGG)_3$ DNA sites. Two different length proteins, $TRF1_{371–439}$ and $TRF1_{337–439}$, were used and the DNA concentration was 4×10^{-8} M in all reactions. Binding to the two-repeat site is shown in lanes 1–6 and to the three-repeat sequence in lanes 7–14. The protein concentration is 5.8×10^{-8} M in lanes 1–5 and 1.74×10^{-7} M in lanes 7–11. The two proteins were mixed in different molar ratios and incubated at 4°C for 1 h before addition of DNA. The ratios of $TRF1_{371–439}$ to $TRF1_{337–439}$ in lanes 1–5 and lanes 7–11 are 1:0, 0.8:0.2, 0.5:0.5, 0.2:0.8 and 0:1. Lanes 6 and 12 contain no protein. In the schematic representation at the sides of the gel the DNA is represented by a line, $TRF1_{371–439}$ by an open circle and $TRF1_{337–439}$ by a filled circle. In lanes 7 and 8 the presence of a third faint complex is due to non-specific binding of an additional $TRF1_{371–439}$ molecule.

were used in the incubations (Fig. 2, lane 3), it can be concluded that the Myb-like domain of TRF1 binds as a monomer to the two-repeat site and furthermore that the binding site of the TRF1 Myb-like motif is contained within the two-repeat sequence TTAGGGT-TAGGG. To investigate further the DNA binding mode of the TRF1 Myb-like domain the gel retardation analysis described above was repeated using a longer DNA site containing three TTAGGG repeats. Figure 2, lanes 7 and 11, shows formation of two complexes: a fast migrating complex that corresponds to binding of a single protein and a slow migrating one that corresponds to binding of two proteins. Figure 2, lanes 8–10 have an intermediate band which corresponds to binding of one molecule of each length protein, which confirms that two molecules of the Myb-like domain can bind to the three-repeat sequence.

Since previous results suggested that the Myb-like domain of TRF1 was not able to bind to DNA (17), we went on to evaluate its DNA binding affinity. The dissociation constant for binding of $TRF1_{371–439}$ to the two-repeat sequence was estimated from a quantitative gel retardation analysis (Fig. 3a) as described in Materials and Methods. The plot in Figure 3b shows the concentration of complex plotted against the concentration of complex divided by the concentration of free DNA. The dissociation constant was estimated from the slope to be $3.2 \pm 0.5 \times 10^{-9}$ M. The concentration of active protein can be calculated from the ordinate intercept and the total protein concentration (see Materials and Methods), which in this case gives a value of 0.28. The activity coefficient corresponding to ~28% active protein may indicate only partial refolding of the protein after the purification step involving a reverse phase column. Next, in order to show that the affinity is specific for the human telomeric repeat (TTAGGG), we performed a competition binding experiment (as described in Materials and Methods) with an oligonucleotide containing two point mutations in the repeat sequence. Figure 3d shows a plot derived from the data presented in Figure 3c. In this plot the concentration of total mutated DNA divided by the concentration of total specific DNA is plotted against the amount of complex without competitor minus the

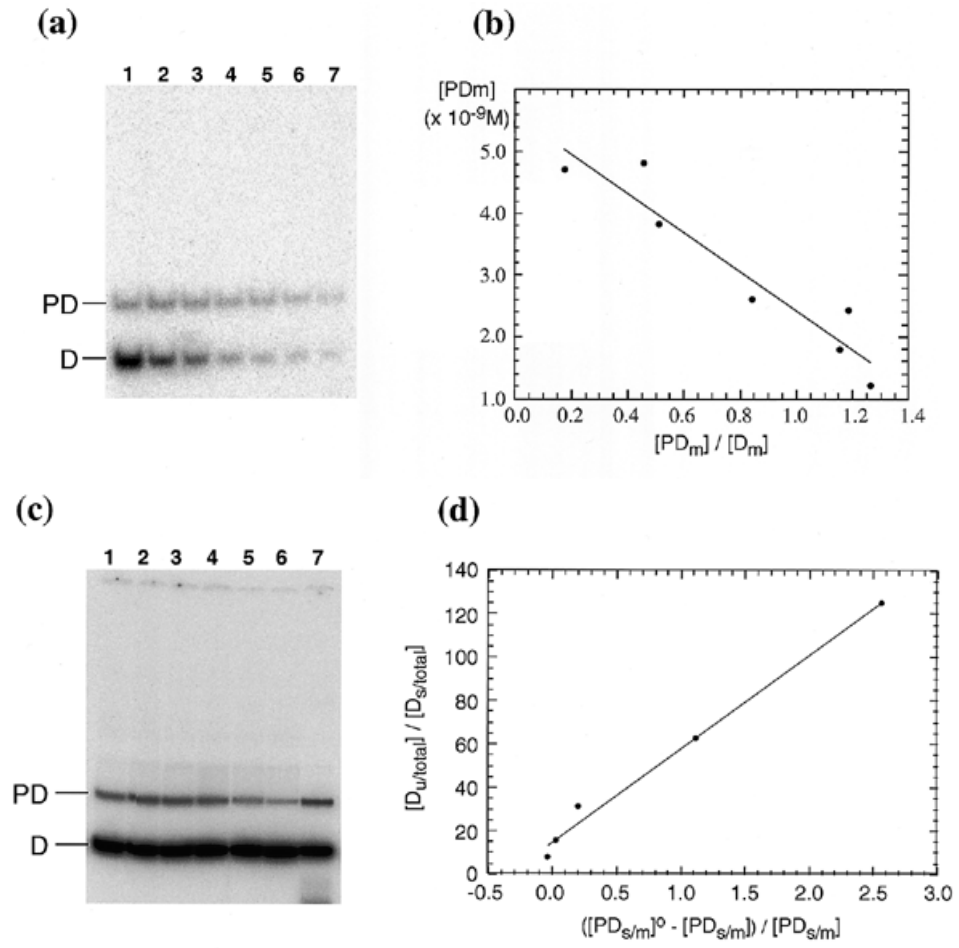

Figure 3. Quantitative binding analysis of the TRF1 Myb-like domain. (**a**) Binding of $TRF1_{371–439}$ to the two-repeat DNA site. The protein concentration is 2×10^{-8} M and the DNA concentrations in lanes 1–7 are 2.95×10^{-8}, 1.7×10^{-8}, 1.1×10^{-8}, 0.70×10^{-8}, 0.50×10^{-8}, 0.32×10^{-8} and 0.18×10^{-8} M. (**b**) Plot of bound DNA $[PD_m]$ versus $[PD_m]/[D_m]$ using data from (a). The dissociation constant K_d was estimated from the slope and the concentration of active protein as the ordinate intercept of the linear regression line (see also Materials and Methods). (**c**) Competition binding experiment with the mutated telomeric DNA site. The protein concentration is 1.4×10^{-8} M and the specific DNA concentration is 1.6×10^{-7} M. The concentrations of the mutated competitor DNA in lanes 1–7 are 0.625×10^{-6}, 0.125×10^{-5}, 0.25×10^{-5}, 0.5×10^{-5}, 1×10^{-5}, 2×10^{-5} and 0 M. (**d**) Plot of $[D_{u/total}]/[D_{s/total}]$ versus $([Pd_{s/m}]^\circ - [PD_{s/m}])/[PD_{s/m}]$ using data from (c). The data point of lane 1 was omitted. The relative dissociation constant K_u/K_s was determined from the slope of the linear regression line (see also Materials and Methods).

amount of complex with competitor divided by the amount of complex with competitor. The slope of this plot gives the ratio of the binding affinity for the mutated DNA versus the specific. The introduction of the point mutations (TTAGG**C**TTAGG**C**) reduced the binding affinity by a factor of 43 ± 3, which is consistent with the results of similar experiments with full-length TRF1 (22). From competition binding experiments using poly(dI·dC) we estimate a non-specific binding constant of $3.8 \pm 0.6 \times 10^{-6}$ M (not shown). The DNA binding affinities of two TRF1 Myb-like domains interacting simultaneously with the three-repeat sequence are essentially identical (not shown), indicating that the proteins bind without positive or negative cooperativity to the three-repeat sequence.

These binding experiments show that the Myb-like domain of TRF1 can bind to DNA with high specificity and affinity as a monomer. This result is in contrast to the c-Myb proteins, which require at least two motifs for efficient DNA binding (15,16). The measured DNA binding affinity lies in the range of values reported for various homeodomains that can also bind sequence specifically to DNA as monomers (23–26). Although interaction of the single Myb-like domain of TRF1 with telomeric DNA is specific, the specificity is likely to be significantly increased in a homodimer of the full-length protein (17) and, presumably, the juxtaposition of two Myb-like motifs on telomeric DNA.

The Myb-like domain of TRF1 binds to the sequence TAGGGTTAG

We have used primer extension analysis (27) to define the outer borders of the binding site of the Myb-like domain of TRF1. For this analysis various length binding sites were generated using Sequenase and dideoxy nucleotides as described in Materials and Methods. These different length binding sites were then incubated with protein and the bound DNA separated from unbound by gel

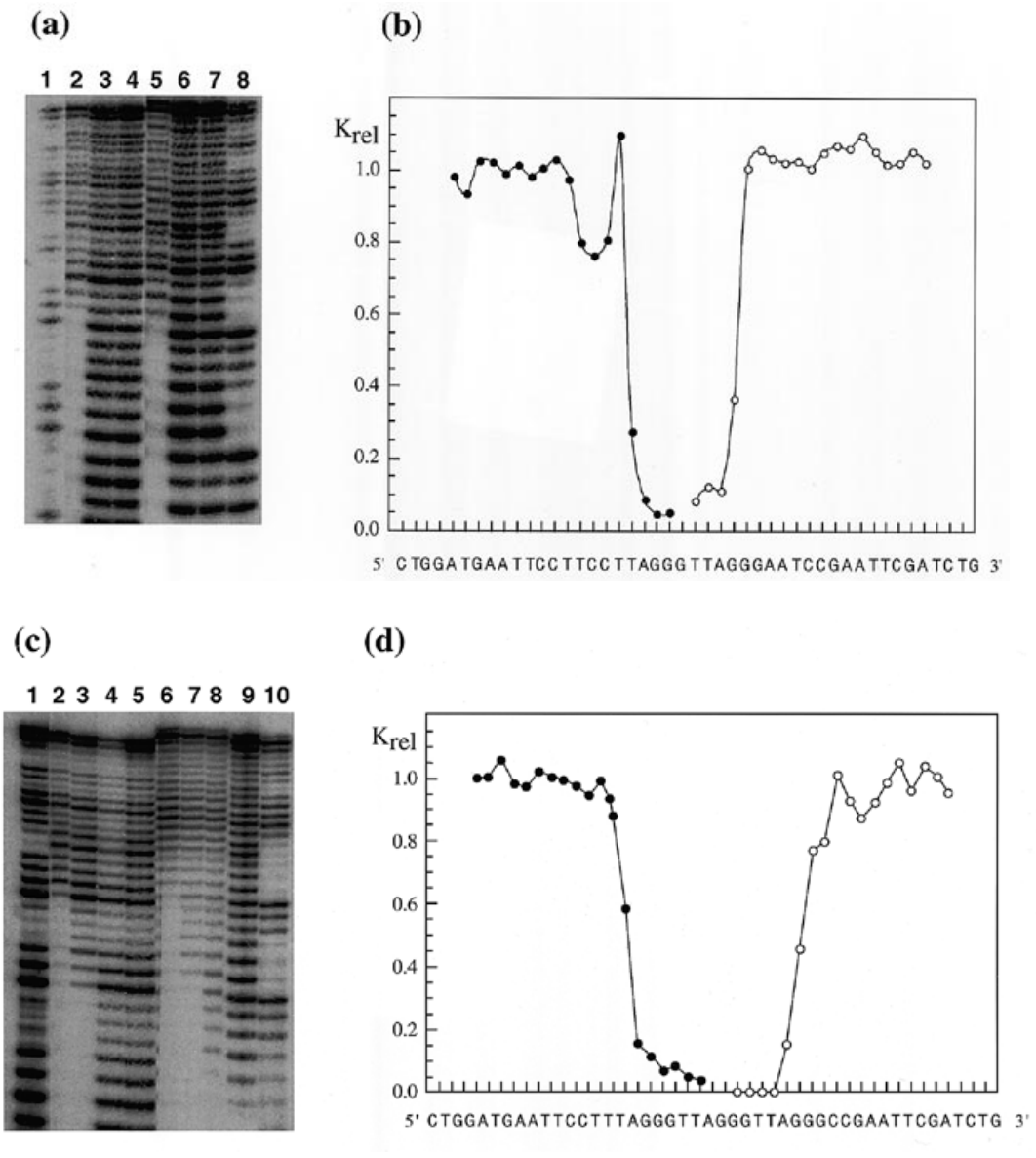

Figure 4. Primer extension analysis of the binding site for one and two of the TRF1 Myb-like domains. Prior to analysis DNA fragments that were bound by $TRF1_{371–439}$ were separated from those that were not bound by gel retardation electrophoresis. Binding reactions were carried out at protein concentrations of 3.2×10^{-8} and 1.3×10^{-8} M for the two- and three-repeat sites respectively. (**a**) Binding to the two-repeat DNA site truncated at the 3′-end of the binding site are shown in lanes 1–4 and those at the 5′-end in lanes 5–8 respectively. Lanes 1 and 8 are marker lanes containing A+T sequencing reactions. Lanes 2 and 5 show bound and lanes 3 and 6 unbound DNA fragments. Lanes 4 and 7 show the total amount of DNA fragments before the binding reaction. (**b**) Plot of relative dissociation constants (K_{rel}) for binding of one Myb-like domain to the two-repeat sequence calculated with data from (a). (**c**) Binding to the three-repeat site. DNA fragments truncated at the 3′-end are shown in lanes 1–5 and those at the 5′-end are shown in lanes 6–10. Lanes 1 and 10 show A+T sequencing reactions. Lanes 2 and 6 show fragments bound by two $TRF1_{371–439}$ molecules and lanes 3 and 7 by one protein molecule. Lanes 4 and 8 contain the unbound DNA fragments and lanes 5 and 10 the total amount of DNA fragments before the binding reaction. (**d**) Plot of K_{rel} for binding of two Myb-like domains on the three-repeat DNA site calculated with data from (c).

retardation electrophoresis. Information on the sequence required for high affinity binding was obtained by comparison of the bound and unbound families of DNA fragments after analysis in denaturing polyacrylamide gels. Figure 4 shows the results of this analysis for binding of the TRF1 Myb-like motif to the two-repeat and three-repeat DNA binding sites. The relative dissociation constant, K_{rel} (for binding of the protein to truncated fragments versus full-length DNA), was plotted for both ends of the relevant DNA binding sites (see Materials and Methods) and is shown in Figure 4b and c. For a single Myb domain bound to the two-repeat sequence the DNA binding site with full binding affinity is 5′-TTAGGGT-TAGG-3′ ($K_{rel} > 0.8$) and that with significant affinity ($K_{rel} > 0.2$) is 5′-TAGGGTTAG-3′ (Fig. 4a and c). In comparison, for two Myb domains bound simultaneously to the three-repeat sequence the binding site with full binding affinity is 5′-TTAGGGTTAGGGTT-AGG-3′ and that with significant affinity is 5′-TAGGGTTAGGGT-TAG-3′ (Fig. 4b and d). Hence, the binding site for two proteins is exactly 6 bp longer than that for one protein. Furthermore, the boundaries of the binding site are the same for two bound proteins on the three-repeat as for a single protein bound to the two-repeat

sequence. This, together with the observation that the DNA binding affinities for one and two proteins are essentially the same (see above), provides experimental evidence that the Myb-like motif of TRF1 recognizes the same sequence within the two-repeat and in the three-repeat sequences. In conclusion, these results are consistent with the TRF1 Myb-like motifs binding in a tandem orientation every 6 bp along telomeric DNA, which corresponds to the length of the human telomere repeat.

Binding of two TRF1 Myb-like motifs with a centre-to-centre spacing of 6 bp is clearly different to the arrangement of the two Myb domains of c-Myb on DNA. In the NMR structure of the c-Myb complex with a specific DNA site (19) the two intramolecularly linked domains contact, in a cooperative manner, binding sites of 3 and 4 bp respectively and together specify a binding site of 6 bp. Therefore, despite a significant sequence similarity between the Myb-like domains of TRF1 and Myb proteins (Fig. 1b), the size of their DNA binding sites and the spacing between adjacent binding domains is different.

Model of the Myb-like motif of TRF1 in complex with DNA

The significant sequence similarity between the DNA binding motifs of the Myb family of transcription regulators and TRF1 (Fig. 1b) suggests that they have a similar three-dimensional structure. We have made use of the available structural information on the DNA binding domain of c-Myb (19), homeodomains (28) and the DNA binding domain of RAP1 (18) in complex with their respective DNA binding sites to build a three-dimensional model of the Myb-like motif of TRF1 and its possible interactions with human telomeric DNA. The second motif of c-Myb (residues 93–135), which shares a sequence identity of 32% with the Myb-like motif of TRF1 (residues 381–425) (Fig. 1b), was used to model the three-helix bundle of TRF1. The TRF1 Myb-like motif can be folded into the three-dimensional structure of the second c-Myb motif without any significant steric clashes (see Materials and Methods). The sequence similarity between c-Myb and TRF1 arises primarily from conservation of structurally important residues that form the hydrophobic core of the three-helix bundles, as well as conservation of basic residues in the DNA recognition helix of these domains (Fig. 1b).

Furthermore, we assumed that the TRF1 Myb-like domain, by analogy with homeodomains, is preceded by a short N-terminal arm. This was based on the following observations. Firstly, the amino acid sequence alignment of the TRF1 Myb-like domain with the Engrailed homeodomain shows that residues in the N-terminal arm of Engrailed that are used to contact bases in the minor groove of DNA are also present in identical positions in TRF1 (residues 378–380) (Fig. 1b). Secondly, in contrast to the DNA binding motifs of c-Myb (19), homeodomains and the Myb-like domain of TRF1 can bind to DNA as monomers with high affinity, presumably due to the additional contacts made by the N-terminal arm (29). Thirdly, the high resolution structure of the telomere binding protein RAP1 shows that each of the two subdomains in addition to the three-helix bundle uses a short N-terminal arm for sequence-specific DNA recognition (18).

Docking of the resulting model for the TRF1 Myb-like domain ($TRF1_{371–439}$) onto telomeric DNA (in B-DNA conformation) was made taking into account the DNA contacts made by the N-terminal arm of Engrailed (28) and the three-helix bundles of the RAP1 subdomains (18). The best fit was obtained when the model of the TRF1 Myb-like domain was aligned over the sequence GGGTTA/TAACCC (Fig. 5). This alignment has a surprising complementarity in the pattern of specific contacts and predicts specific contacts from both the N-terminal arm and the three-helix bundle. Residues R378 and R380 in the N-terminal arm of the TRF1 Myb-like domain could interact with two bases within the telomeric repeat GGGT**T**A/**T**AACCC in a similar way to the corresponding residues in the Engrailed homeodomain, which contact two bases in the binding site TAAT**T**A/**T**AATTA (28). A similar contact is made by a residue in the N-terminal arm of each of the RAP1 subdomains (GGGTGT/**A**CACCC). This interaction positions the three-helix bundle of TRF1 in the major groove over the sequence **GGGT**TA/TA**ACCC**. Significantly, a very similar telomeric sequence **GGGT**GT/AC**ACCC** is recognized by the three-helix bundles of the two RAP1 domains. The analogy between TRF1 and RAP1 is further supported by the observation that key residues for specific interaction are present at corresponding positions on the DNA recognition helices of TRF1 and RAP1 (RAP1, 541, W**RD**RF**R**; TRF1, 420, L**KD**RW**R**) (Fig. 1b). Hence TRF1 and RAP1 could make a similar pattern of protein–DNA contacts, e.g. residues K421 and R425 of TRF1 are positioned to interact with bases in the G base triplet (**GGG**TTA/

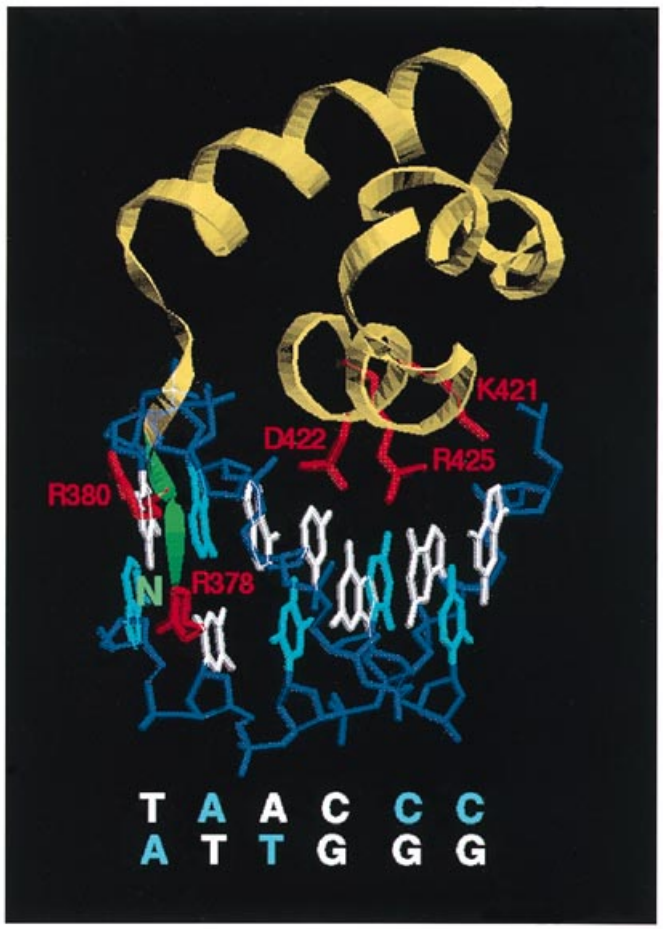

Figure 5. Model of the TRF1 Myb-like domain–DNA complex. The domain is aligned on the sequence GGGTTA/TAACCC. The N-terminal arm is shown in green and the three-helix bundle in yellow. Amino acid side chains predicted to make specific interactions are shown in red and DNA bases predicted to be contacted are shown in white.

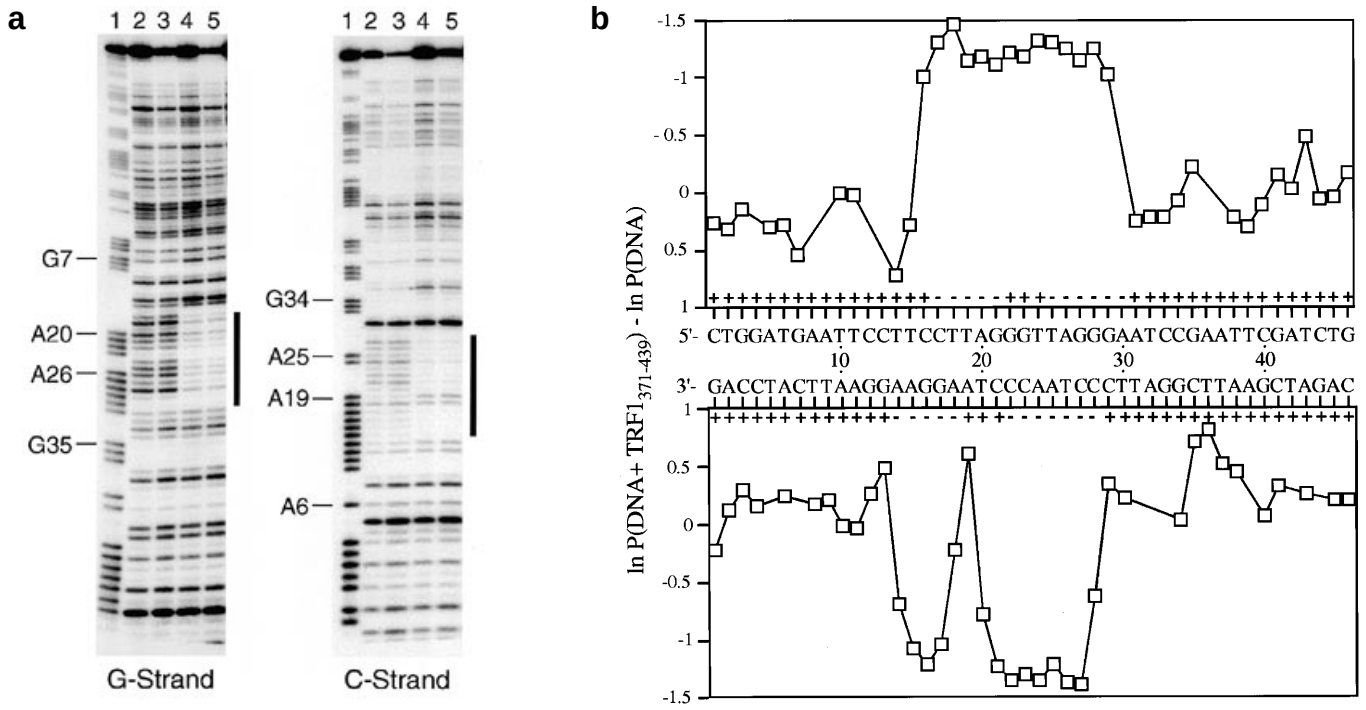

Figure 6. DNase I footprint of the TRF1 Myb-like domain bound to the two-repeat DNA site. Binding reactions were carried out at a DNA concentration of 1×10^{-9} M and $TRF1_{371–439}$ at 4×10^{-8} M. For each strand lanes 2 and 3 show the cleavage pattern of the naked DNA and lanes 4 and 5 that of the protein–DNA complex. Lane 1 is a marker tract containing G+A Maxam–Gilbert sequencing reactions. The region protected by the protein from cleavage is indicated. (**b**) Experimental and theoretical DNase I footprint. The experimental difference probability plot was calculated using data from the gel shown in (a). Negative numbers show the region of the DNA protected from DNase I cleavage. The theoretical DNase I footprint is aligned above the sequence. Bonds predicted to be protected are indicated by – and those accessible to cleavage by +.

TAACCC) and D422 of TRF1 could contact either an adenine or cytosine (GGGTTA/TA**AC**CC). Further discussion about the detailed pattern of contacts would be speculative, since they are likely to be affected by other factors such as water-mediated protein–DNA interactions and the exact orientation of the DNA recognition helix within the major groove, which is different for c-Myb and RAP1 (18,19).

In addition to base-specific contacts to the sequence GGGTTA, the model described above (Fig. 5) predicts contacts with the ribose–phosphate backbone. When such contacts are taken into consideration the binding site of the TRF1 Myb-like domain extends over the sequence AGGGTTA. Thus the binding site predicted from the model is in very good agreement with the binding site TAGGGTTAG defined from the primer extension analysis. The experimentally determined binding site is longer than that obtained from the model by one base pair at each end, but this is probably due to end effects of the method, rather than reflecting real contacts.

Correlation between experimental and theoretical DNase I footprints

The validity of the alignment of the Myb-like domain of TRF1 on the human telomeric repeats was tested further by comparing the experimentally determined DNase I protection pattern with a theoretical one derived from the model described above (Fig. 6). The theoretical DNase I footprint was created using the three-dimensional structure of the DNase I–DNA complex to obtain precise information on the relationship between the binding and cleavage sites and steric exclusion (30,31). The area protected from DNase I cleavage by bound protein on each of the two DNA strands was determined by making a series of superpositions of the DNase I structure over the model of the TRF1 Myb-like domain–DNA complex (see Materials and Methods).

The experimental DNase I footprint for both the C-rich and G-rich strands spans ~11 bp (Fig. 6a and b) encompassing the sequence GGGTTA/TAACCC that is predicted to be specifically bound in the model of the protein–DNA complex (Fig. 5). For the C-rich strand the theoretical and experimental DNase I footprints are in perfect agreement (Fig. 6b). The model predicts that the protection between A_{19} and C_{29} arises from binding of the three-helix bundle and the N-terminal arm that crosses from the major groove into the minor groove at A_{25}. The hypersensititve site A_{19} represents the left-hand border of the binding site of the protein. From the model it can be seen that a second protected region on the C-rich strand (A_{15}–A_{18}) arises from steric hindrance. For the G-rich strand the observed and predicted footprints are also in good overall agreement, except for a short protected region (G_{22}–T_{24}), which from the model we would predict to be accessible to DNase I cleavage (Fig. 6b). This difference might be due to the presence of an additional nine amino acids that are not included in the model (Fig. 5). This region extends beyond the N-terminal arm and might be in a position to protect G_{22}–T_{24}. An alternative explanation is that the protein may induce a conformational change in the DNA, preventing DNase I cleavage.

In summary, the good agreement we have found between the theoretical and experimental DNase I footprints (Fig. 6) provides additional evidence for specific binding of the TRF1 Myb-like domain to the human telomeric sequence GGGTTA.

DISCUSSION

We show, using gel retardation analysis (Figs 2 and 3), primer extension (Fig. 4) and DNase I footprints analysis (Fig. 6), that the isolated Myb-like domain of the human telomere binding factor TRF1 binds sequence specifically and with significant affinity to the sequence AGGGTTA. Furthermore, the size of this binding site, 6–7 bp, is more consistent with the Myb-like domain of TRF1 binding to DNA in the same way as a homeodomain rather than a Myb domain. Based on amino acid sequence similarities and precise three-dimensional information from the structurally related DNA binding domains of c-Myb, the yeast telomere binding protein RAP1 and the Engrailed homeodomain, we propose a three-dimensional model for interaction of the Myb-like domain of TRF1 with its telomeric recognition site (Fig. 5). The model predicts that the TRF1 Myb-like domain consists of a three-helix bundle similar to that seen in the NMR structure of c-Myb (19) and a short N-terminal arm of the type seen in homeodomains (see for example 28) and the two DNA binding domains of RAP1 (18). The N-terminal arm extends DNA recognition from 3–4 bp, as seen for the c-Myb domains, to 5–6 bp, as seen for the RAP1 domains. One of the most interesting results emerging from the model of the human TRF1–DNA complex is that base-specific contacts are made within the sequence GGGTTA, which is very similar to the (G)GGTGT sequence recognized by each of the two domains of RAP1. Recognition of similar binding sites for RAP1 and TRF1 arises from conservation in the two proteins of the most important DNA binding residues, resulting in a conserved pattern of protein–DNA contacts.

The telomeric sequence motif GGGTTA is not only conserved in vertebrates but is also found in the telomeres of more distantly related eukaryotes, e.g. *Arabidopsis* and *Trypanosomes*. A partial subsite, the sequence GGTTA, forms part of the most frequently occurring repeating unit GGTTAC of the rather irregular telomeric DNA of *S.pombe*. The *S.pombe* telomere binding factor TAZ1 contains a Myb-like domain (7) which shows significant sequence similarity (30% identity) and hence is expected to have the same structure as the Myb-like domain of human TRF1 (8). The sequence similarity includes a set of basic residues within the region expected to be the DNA recognition helix (4,9) as well as in the region proposed to form an N-terminal arm in TRF1. Although the binding site of TAZ1 on *S.pombe* telomeric DNA is not known, conservation in the protein sequence of the Myb-like domains of TAZ1 and TRF1 is consistent with recognition of the sequence GGTTA. In summary, the available information on telomere repeat binding proteins indicates that they use a common protein fold to interact with similar sequences present in the telomeric repeats of eukaryotes.

The tandem organization of short repeats results in long arrays of double-stranded DNA containing a number of potentially closely spaced binding sites. The binding studies presented here not only show that an isolated TRF1 Myb-like domain binds to the sequence GGGTTA, but that two domains can bind next to each other with a centre-to-centre spacing of 6 bp (Figs 1 and 3). In the model in Figure 7 two Myb-like domains of TRF1 have been docked onto telomeric DNA with a 6 bp centre-to-centre

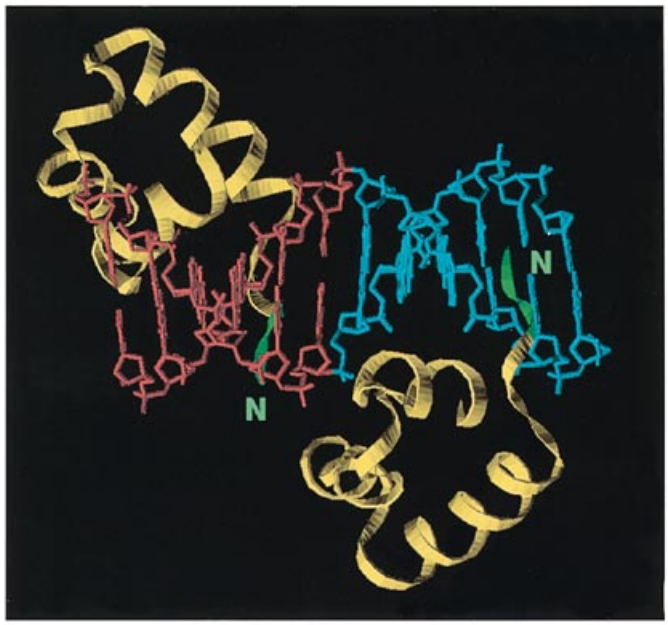

Figure 7. Model for binding of two TRF1 Myb-like domains. The domains are spaced 6 bp apart. The N-termini of the domains are indicated.

spacing. The two domains can interact with the closely spaced telomeric repeats with no steric clashes, since the consequence of this spacing is that adjacent domains are bound on opposite faces of the DNA double helix (Fig. 7). This observation is relevant to the question of how the TRF1 homodimer binds to telomeric repeats. Full-length TRF1 binds to DNA as a preformed homodimer and both Myb-like domains are apparently required for high affinity binding, providing experimental evidence that both Myb-like domains are involved in DNA recognition (17). The observation that TRF1 binds weakly to a three-repeat sequence and with considerably higher affinity to 6 and 12 TTAGGG repeats (17,22) suggests, however, that binding of dimeric full-length TRF1 may be more complex than binding of two isolated Myb-like domains. The use of two Myb-like domains to recognize DNA provides a means of increasing both the binding affinity and specificity, by allowing the distance between the two binding sites, or half-sites, to be recognized. The linking of two domains is also seen in RAP1, but in this case they are intramolecularly linked and the centre-to-centre spacing is 8 bp (18).

Further biochemical studies will be required to reveal the DNA binding mode of dimeric TRF1. In addition, structural studies on a number of Myb-like domains from telomeric proteins from different organisms in complex with their respective telomeric DNA binding sites should reveal to what degree the patterns of protein–DNA interactions at telomeres are evolutionarily conserved.

ACKNOWLEDGEMENTS

We are indebted to Titia de Lange for the clone of TRF1, discussions and comments on the manuscript. We thank Kate Sparks for N-terminal amino acid sequencing and amino acid analysis, Song Tan and Timothy Richmond for the primer extension protocol and Lynda Chapman and Helena Taylor for

comments on the manuscript. P.K. also acknowledges support by CEC grant CHRX-CT92-0022.

REFERENCES

1 Zakian,V.A. (1995) *Science*, **270**, 1601–1607.
2 Sandell,L.L. and Zakian,V.A. (1993) *Cell*, **75**, 729–739.
3 Harley,C.B. (1995) In Blackburn,E.H., and Greider,C.W. (eds), *Telomeres*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 247–263.
4 König,P. and Rhodes,D. (1997) *Trends Biochem. Sci*., **22**, 43–47.
5 Berman,J., Tachibana,C. and Tye,B.-K. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 3713–3717.
6 Larson,G.P., Castanotto,D., Rossi,J.J. and Malafa,M.P. (1994) *Gene*, **150**, 35–41.
7 Cooper,J.P., Nimmo,E.R., Allshire,R.C. and Cech,T.R. (1997) *Nature*, **385**, 744–747.
8 Chong,L., van Steensel,B., Broccoli,D., Erdjument-Bromage,H., Hanish,J., Tempst,P. and de Lange,T. (1995) *Science*, **270**, 1663–1667.
9 Bilaud,T., Koering,C.E., Binet-Brasselet,E., Ancelin,K., Pollice,A., Gasser,S.M. and Gilson,E. (1996) *Nucleic Acids Res*., **24**, 1294–1303.
10 Broccoli,D., Chong,L., Oelmann,S., Fernald,A.A., Marziliano,N., van Steensel,B., Kipling,D., Le Beau,M.M. and de Lange,T. (1997) *Hum. Mol. Genet*., **6**, 69–76.
11 Broccoli,D., Smogorzewska,A., Chong,L. and de Lange,T. (1997) *Nature Genet*., **17**, 231–235.
12 Kyrion,G., Boakye,K.A. and Lustig,A.J. (1992) *Mol. Cell. Biol*., **12**, 5159–5173.
13 Krauskopf,A. and Blackburn,E.H. (1996) *Nature*, **383**, 354–357.
14 van Steensel,B. and de Lange,T. (1997) *Nature*, **385**, 740–743.
15 Howe,K.M., Reakes,C.F.L. and Watson,R.J. (1990) *EMBO J*., **9**, 161–169.
16 Tanikawa,J., Yasukawa,T., Enari,M., Ogata,K., Nishimura,Y., Ishii,S. and Sarai,A. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 9320–9324.
17 Bianchi,A., Smith,S., Chong,L., Elias,P. and de Lange,T. (1997) *EMBO J*., **16**, 1785–1794.
18 König,P., Giraldo,R., Chapman,L. and Rhodes,D. (1996) *Cell*, **85**, 125–136.
19 Ogata,K., Morikawa,S., Nakamura,H., Sekikawa,A., Inoue,T., Kanai,H., Sarai,A., Ishii,S. and Nishimura,Y. (1994) *Cell*, **79**, 639–648.
20 Pabo,C.O. and Sauer,R.T. (1992) *Annu. Rev. Biochem*., **61**, 1053–1095.
21 Gehring,W.J., Affolter,M. and Bürglin,T. (1994) *Annu. Rev. Biochem*., **63**, 487–526.
22 Zhong,Z., Shiue,L., Kaplan,S. and De Lange,T. (1992) *Mol. Cell. Biol*., **12**, 4834–4843.
23 Affolter,M., Percival-Smith,A., Muller,M., Leupin,W. and Gehring,W.J. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 4093–4097.
24 Florence,B., Handrow,R. and Laughon,A. (1991) *Mol. Cell. Biol*., **11**, 3613–3623.
25 Ades,S.E. and Sauer,R.T. (1994) *Biochemistry*, **33**, 9187–9194.
26 Carra,J.H. and Privalov,P.L. (1997) *Biochemistry*, **36**, 526–535.
27 Liu-Johnson,H.-N., Gartenberg,M.R. and Crothers,D.M. (1986) *Cell*, **47**, 995–1005.
28 Kissinger,C.R., Liu,B.S., Martin-Blanco,E., Kornberg,T.B. and Pabo,C.O. (1990) *Cell*, **63**, 579–590.
29 Ades,S.E. and Sauer,R.T. (1995) *Biochemistry*, **34**, 14601–14608.
30 Weston,S.A., Lahm,A. and Suck,D. (1992) *J. Mol. Biol*., **226**, 1237–1256.
31 Fairall,L. and Rhodes,D. (1992) *Nucleic Acids Res*., **20**, 4727–4731.
32 Eckstein,F. (1991) *Oligonucleotides and Analogues*. IRL Press, Oxford, UK.
33 Gerchman,S.E., Graziano,V. and Ramakrishnan,V. (1994) *Protein Expression Purificat*., **5**, 242–251.
34 Dubendorff,J.W. and Studier,F.W. (1991) *J. Mol. Biol*., **219**, 45–59.
35 Riggs,A.D., Suzuki,H. and Bourgeoss,S. (1970) *J. Mol. Biol*., **48**, 67–83.
36 Kim,J., Zwieb,C., Wu,C. and Adhya,S. (1989) *Gene*, **85**, 15–23.
37 Smith,J. and Singh,M. (1996) *BioTechniques*, **20**, 1082–1087.
38 Lutter,L.C. (1978) *J. Mol. Biol*., **124**, 391–420.
39 Jones,T.A., Zou,J.-Y. and Cowan,S.W. (1991) *Acta Crystallogr*., **A47**, 110–119.