Nucleotide shuffling and ssDNA recognition in *Oxytricha nova* telomere end-binding protein complexes

Douglas L. Theobald*1* and Steve C. Schultz*1,2*

Department of Chemistry and Biochemistry, University of Colorado at Boulder, Boulder, CO 80309-0215 and *Math/Science Division, Diné College, Tséline, Navajo Nation, AZ 86556, USA

*1*Corresponding authors
e-mail: theobal@colorado.edu and schultz@dinecollege.edu

Sequence-specific protein recognition of single-stranded nucleic acids is critical for many fundamental cellular processes, such as DNA replication, DNA repair, transcription, translation, recombination, apoptosis and telomere maintenance. To explore the mechanisms of sequence-specific ssDNA recognition, we determined the crystal structures of 10 different non-cognate ssDNAs complexed with the *Oxytricha nova* telomere end-binding protein (OnTEBP) and evaluated their corresponding binding affinities (PDB ID codes 1PH1–1PH9 and 1PHI). The thermodynamic and structural effects of these sequence perturbations could not have been predicted solely upon the cognate structure. OnTEBP accommodates non-cognate nucleotides by both subtle adjustments and surprisingly large structural rearrangements in the ssDNA. In two complexes containing ssDNA intermediates that occur during telomere extension by telomerase, entire nucleotides are expelled from the complex. Concurrently, the sequence register of the ssDNA shifts to re-establish a more cognate-like pattern. This phenomenon, termed nucleotide shuffling, may be of general importance in protein recognition of single-stranded nucleic acids. This set of structural and thermodynamic data highlights a fundamental difference between protein recognition of ssDNA versus dsDNA.

**Keywords:** DNA recognition/protein–DNA interactions/single-stranded DNA/specificity/telomeres

**Introduction**

Specific recognition of single-stranded nucleic acids by proteins is crucial for maintaining, manipulating and utilizing the genetic material contained in cells. Single-stranded nucleic acid sequences participate in a large variety of fundamental dynamic biological functions, including telomere regulation, mRNA splicing, DNA damage repair, DNA replication, recombination, chromosomal condensation, transcription regulation, translation, apoptosis, stages in bacterial and viral life cycles, and retrotransposition (Hohjoh and Singer, 1997; Miller et al., 1997; Waga and Stillman, 1998; Handa et al., 1999; Deo et al., 1999; Gallia et al., 2000; Tateishi et al., 2000; Changela et al., 2001; Gasior et al., 2001; Gomis-Ruth et al., 2001; Negroni and Buc, 2001; Pennock et al., 2001). Consequently, protein recognition of single-stranded nucleic acids has been clearly implicated in myriad pathological processes in humans, such as cancer, aging, various infectious diseases, autoimmune illnesses and drug dependency (Vaziri et al., 1993; Osugi et al., 1996; Bodnar et al., 1998; Stevens and Glick, 1999).

In the past few years, several high-resolution structures of sequence-specific ssDNA and ssRNA complexes have been determined: *Dm* sex-lethal, *HnUP1*, *Bac*TRAP, *EcoRho* transcription terminator, *Oxytricha nova* telomere end-binding protein (OnTEBP), *HnSF1*, *Hn* polyA-binding protein, FBP and the tumor suppressor BRCA2 (Antson et al., 1999; Bogden et al., 1999; Deo et al., 1999; Ding et al., 1999; Handa et al., 1999; Classen et al., 2001; Horvath and Schultz, 2001; Liu et al., 2001; Braddock et al., 2002; Peersen et al., 2002; Yang et al., 2002). The interactions observed in these structures hint at possible mechanisms for sequence-specific recognition, and several common themes have emerged (Theobald et al., 2003a). Single-stranded nucleic acids are bound on the surfaces of β-sheets. Phosphodiester groups are extensively solvent exposed, while DNA or RNA bases are in intimate contact with the protein. Bases are typically bound by stacking with aromatic side chains, and hydrogen bond donor and acceptor groups provide recognition of the edges of the bases.

In general, the molecular basis of sequence specificity has been inferred from the structure of the cognate complex. However, structural comparisons of different cognate and non-cognate complexes are required for elucidating the mechanisms of preferential binding, while parallel thermodynamic studies are necessary to evaluate the energetic importance of observed interactions (von Hippel, 1994). Accordingly, we have investigated the molecular basis of cognate ssDNA recognition via comparisons of 10 different non-cognate OnTEBP complexes, using both crystallographic studies and thermodynamic measurements (PDB ID codes 1PH1–1PH9 and 1PHI). Two of these non-cognate ssDNAs correspond to DNA intermediates which occur transiently during extension of the telomere by the enzyme telomerase (Tel+T and Tel-3 in Table I) (Zahler and Prescott, 1988). This structural work, complemented by relevant energetic analyses, provides the most detailed information available revealing how any sequence-specific ssDNA–protein complex responds to non-cognate sequences.

OnTEBP recognizes a 16-nucleotide 3’-terminal single-stranded *T₅G₅T₅* telomeric DNA with high affinity (Raghuraman and Cech, 1990; Fang et al., 1993) and strong sequence specificity (Price and Cech, 1987; Gray, 1992; Classen et al., 2003) (Figure 1). OnTEBP contains two protein subunits, a 56 kDa α-subunit and a 41 kDa β-subunit, which together form a remarkably stable α:β:ssDNA ternary complex with 1:1:1 stoichiometry. In
the structure of the cognate complex the ssDNA is bound in a cleft, formed from the surfaces of β-sheets presented by multiple OB-folds, between the N-terminal domain of α and the globular portion of β (Horvath and Schultz, 2001). The DNA is folded into a contorted configuration in which generally the bases are enclosed within the complex and the phosphodiester groups are solvent exposed. The 3’-terminal deoxyribose is recognized via burial within the ternary complex, a useful feature for these studies as it positions the DNA in a specific register.

**Sequence specificity and binding affinity of non-cognate sequences**

The high degree of specificity exhibited by OnTEBP binding to its cognate sequence was confirmed using a competition assay (Sidorova and Rau, 1996). The assay directly measures $K_{\text{spec}}$, which is the ratio of the affinity constant of the protein for cognate DNA to the affinity constant of the protein for non-specific DNA. Previous work has shown that various non-specific ssDNAs give an $K_{\text{spec}} \approx 10^5$ for the OnTEBP ternary complex (Gray, 1992), and we determined that $K_{\text{spec}} = 8.5 \times 10^4$ for a cognate sequence relative to a pool of random ssDNA of the same length. Thus the OnTEBP complex is highly sequence specific, with a $K_{\text{spec}}$ comparable with that of classical double-stranded DNA binding proteins such as EcoRI (Lesser et al., 1990), BamHI (Engler, 1998), lac repressor (Mossing and Record, 1985) and Cro protein (Takeda et al., 1992).

All 10 of the non-cognate ssDNAs bind with significantly weaker affinity than the cognate sequence, with the single exception of OxyG10I (Table II). Relative to the cognate sequence, the single base substitutions and modifications result in increases in $K_D$ of up to a factor of 100. A $K_{\text{rel}}$ of 6700 [$K_{\text{rel}} = K_D(\text{non-cognate})/K_D(\text{cognate})$] is observed for a ssDNA in which the three 3’-most cognate nucleotides are missing (OxyTel-3), equivalent to an average increase in $K_D$ of 19-fold per missing nucleotide. To elucidate the molecular basis of the energetic perturbations seen with these non-cognate ssDNAs, we solved the crystal structures of the 10 corresponding non-cognate complexes.

---

**Table I. ssDNA sequences**

<table>
<thead>
<tr>
<th>ssDNA name</th>
<th>ssDNA sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tel+T</td>
<td>GGGTGTGTTGCGGTT</td>
</tr>
<tr>
<td>Tel-3</td>
<td>GGGTTTTT</td>
</tr>
<tr>
<td>G3X</td>
<td>GG-GTGGTGGGG</td>
</tr>
<tr>
<td>G10I</td>
<td>GGGTTTTTGGGG</td>
</tr>
<tr>
<td>G10X</td>
<td>GGGTTTTTGGGG</td>
</tr>
<tr>
<td>G10A</td>
<td>GGGTTTTTGGGG</td>
</tr>
<tr>
<td>G10T</td>
<td>GGGTTTTTGGGG</td>
</tr>
<tr>
<td>G11C</td>
<td>GGGTTTTTGGGG</td>
</tr>
<tr>
<td>Tel (cognate)</td>
<td>GGGTTTTTGGGG</td>
</tr>
<tr>
<td>TelRNDSM</td>
<td>NNNNNNNNNNN</td>
</tr>
</tbody>
</table>

All oligonucleotides are numbered beginning at one from the 5’-end. ‘-’ = abasic; ‘N’ = a random mixture of all four nucleotides.
complex. The other seven non-cognate complexes display interesting subtle structural changes illustrating how a protein can accommodate various non-cognate nucleotide substitutions.

**Register shifting and nucleotide shuffling in non-cognate complexes**

*Addition of a thymine nucleotide at the 3' end of the cognate ssDNA (Tel+T complex) (K<sub>rel</sub> = 5.3).* The 3'-OH of the cognate ssDNA is buried deep within the complex and hydrogen bonds to the backbone amide of αK66 (Figure 2A). There is no apparent room for an additional nucleotide. Therefore it was surprising to discover that Tel+T, a ssDNA with a single thymine nucleotide added to the 3' end of the cognate sequence, binds with nearly the same affinity as the cognate ssDNA. Notably, Tel+T ssDNA occurs as an intermediate during polymerization of the telomeric DNA by telomerase (Zahler and Prescott, 1988).

The structure of the Tel+T complex reveals that *On*TEBP binds the 3'-OH of Tel+T in precisely the same protein site as does the cognate complex (Figure 2B). The DNA register shifts one base towards the 5' end compared with the cognate sequence in order to maintain this 3'-OH interaction. However, this register shift is not propagated throughout the entire length of the DNA. Rather, T8 flips entirely out of contact with the protein and into solution. This nucleotide expulsion arrests the register shift initiated by the 'extra' thymidine (T13) and re-establishes the proper cognate register for nucleotides G1–T7.

The shift in register of the 3'-loop region effectively results in two base substitutions. G12 and T8 of the cognate sequence are replaced with T13 and G9, respectively, of Tel+T. Therefore the T8:G12 stack that closes the 3' loop in the cognate complex is exchanged for a G9:T13 stack in the Tel+T complex. Otherwise, the molecular conformations of the cognate and non-cognate complexes are nearly identical, including the portions of the DNA backbone that flank the flipped-out nucleotide Tel+T T8 (Figure 2C). All aromatic stacking interactions are maintained.

The Tel+T non-cognate ssDNA results in a strikingly small reduction in binding energy relative to the cognate ssDNA (ΔG<sub>rel</sub> = 4.2 kJ/mol). Though the Tel+T ssDNA has undergone a register shift involving six nucleotides, two effective nucleotide substitutions and a dramatic repositioning of the ssDNA backbone, the complex maintains nearly all cognate ssDNA–protein interactions. If the crystal structure of the non-cognate Tel+T complex had been unavailable, binding results for the Tel+T DNA might have been misinterpreted as a lack of 3'–end binding rather than shuffling of nucleotides.
Structure of an OnTEBP complex with three 3’-nucleotides removed (Tel-3) ($K_{rel} = 6700$). We were similarly surprised to obtain crystals of a complex containing a DNA in which three nucleotides at the 3’ end were entirely removed (Tel-3). Similar to Tel+T, the Tel-3 ssDNA corresponds to a telomeric intermediate that occurs during extension of the chromosomal terminus by telomerase. As shown in Figure 3B, OnTEBP binds to the Tel-3 ssDNA by shifting the DNA register one nucleotide towards the 3’ end relative to the cognate sequence, leaving the two 3’-most nucleotide binding sites entirely unoccupied. This register shift, unlike that seen in Tel+T and G3X, is propagated for the entire length of the DNA sequence (involving a total of nine nucleotides). This results in base substitutions at two positions: G4 in the Tel-3 complex replaces cognate T5, and T8 in the Tel-3 complex replaces cognate G9.

Three nucleotide binding sites, corresponding to T8, G11 and G12 of the cognate complex, are completely unoccupied in the Tel-3 structure, yet the protein side-chain conformations are very similar to those in the cognate complex. If the Tel-3 ssDNA had not shifted towards the 3’ end by one nucleotide, the cognate G10 binding pocket would also have remained vacant. Nevertheless, the register shift observed in the Tel-3 complex allows G9 to occupy the cognate G10 binding site.

As one might predict, Tel-3 results in the largest energetic destabilization of the non-cognate ssDNAs ($\Delta \Delta G_{rel} = 22 \text{ kJ/mol}$, or $\sim 7 \text{ kJ/mol}$ per removed nucleotide). The DNA–protein contacts found in the G10 binding site appear strong enough to outweigh the introduction of non-cognate interactions at the T5 and G9 binding sites arising from the shift of the ssDNA. Relative to the cognate complex, the Tel-3 complex has lost 352 Å² of contact surface area between the protein and ssDNA, amounting to nearly a quarter of the contact surface area in the cognate complex. Presumably, the
majority of the $\Delta \Delta G_{rel}$ for the Tel-3 complex arises from the evacuation of the predominantly non-polar T8, G11 and G12 binding sites.

Abasic G3X nucleotide substitution at position G3 ($K_{rel} = 56$). Substitution of G3 with an abasic nucleotide also results in a large unanticipated change in the ssDNA conformation (Figures 4B and C). The abasic nucleotide is entirely expelled from the complex, shifting the register of the ssDNA in the 3' direction. G1 and G2 occupy the protein binding sites that G2 and G3 occupy in the cognate crystal structure. The abasic G3X ribose is flipped out of contact with the $\alpha$-protein subunit and into solution with virtually no repositioning of the adjoining bases. All the interactions observed between the protein and the ssDNA observed in the cognate complex are maintained in the non-cognate G3X complex.

If the non-cognate G3X crystal structure were not available, the most parsimonious explanation for our thermodynamic results would be that the energetic cost of the G3X mutation ($\Delta \Delta G_{rel} = 10$ kJ/mol) is due to loss of the various protein contacts with the G3 base. However, the observed difference in binding affinities does not result from contacts perturbed at the G3 binding site, since all contacts there are identical to the cognate complex. Apparently, $\Delta \Delta G_{rel}$ results from the alternate conformation of the G3X phosphodiester backbone and from protein contacts perturbed at G1. This energetic difference is surprisingly small given the extent of the conformational rearrangement of the G3X ssDNA backbone.

**Interactions with G10**

We chose to investigate G10 because it occupies a pivotal position in the complex, bound between the $\alpha$- and $\beta$-subunits, making an intriguing stacking interaction with the guanidium portion of an arginine side chain (Figure 5A). The G10 base is sandwiched between $\alpha$Y239 and $\beta$R140, the latter of which packs flat against the base and interacts electrostatically with the phosphodiester group on the 5' side of this nucleotide. Two direct hydrogen bonds are made between G10 and the protein: N7 to $\beta$K145, and the exocyclic amino group to the backbone oxygen of $\beta$S102. The exocyclic oxygen participates in a water-mediated interaction with the backbone amide from $\beta$A110 (omitted from Figure 5 for clarity). Interestingly, G10 adopts the syn conformation about the C1'-N glycosyl bond.

**G10I complex ($K_{rel} = 0.5$).** When G10 is replaced with inosine, which lacks only the exocyclic amino group, only slight changes in the complex are observed (Figure 5B). Relative to the cognate complex, the direct hydrogen bond between the backbone carbonyl group of residue $\beta$S102 to the exocyclic amino group of G10 cannot form in the non-cognate G10I complex. All other structural features remain identical. Reflecting the lack of structural rearrangement, the G10I substitution results in a statistically insignificant energetic perturbation. The $\beta$S102-G10:N2 hydrogen bond makes no measurable contribution to specific recognition.
previously made to βK145 from the cognate guanine. In total, the removal of an entire base at position G10 results in the loss of 111 Å² of contact surface area relative to the cognate complex.

As might be expected, the abasic modification results in the largest energetic loss at this position, yet this loss is only 11 kJ/mol. The novel water interaction in the G10X complex, combined with the slight repositioning of two amino acid side chains to favorable rotamers, may partially compensate for the loss of interactions with the guanine base.

G10A complex (K_rel = 35). In the G10A complex, the conformation of the adenine base is flipped nearly 180° about the C1'–N glycosyl bond relative to the cognate guanine, and it sits in an anti conformation (Figure 5D). Despite this 180° rotation of the base, the adenine remains stacked between βR140 and αY239 with essentially no change in the positions of the side chains. Two direct hydrogen bonds are lost (G10:N2 to the βS102 backbone carbonyl and G10:N7 to βK145), and the adenine lacks any direct hydrogen-bond interactions with the protein. An indirect water-mediated interaction with βA110 is maintained in a manner similar to that of the cognate complex.

The G10A substitution results in a loss of 8.8 kJ/mol of binding energy. Guanine is distinguished from the other three DNA bases by its preference for the syn conformation. The rather small loss of energy due to this substitution may result from a combination of altered stacking, loss of hydrogen bond formation and the anti versus syn preference of the adenosine nucleotide.

G10C complex (K_rel = 37). As illustrated in Figure 5E, with G10C a pyrimidine substitutes for a purine with an overall molecular orientation similar to the cognate complex. The N3 of the cytosine hydrogen bonds with βK145, just as the guanine N7 does in the cognate complex. The indirect water-mediated interaction between the cognate guanine O6 carbonyl oxygen and the backbone amide of βA110 is replaced by a similar water-mediated interaction between the exocyclic amino group (N4) of the cytosine and the same backbone amide of βA110. This nucleotide substitution results in only two major differences: (1) the loss of a single direct hydrogen bond at the protein–DNA interface (cognate G10:N2 to the βS102 backbone carbonyl), and (2) the introduction of a small cavity between the ssDNA and the protein where the guanine N1, N2 and O6 are positioned in the cognate complex. The βR140 and αY239 side chains are unchanged and stack flat against the cytosine base.

Like G10A, the G10C substitution results in a loss of ~9.0 kJ/mol of binding energy, likely due to the altered stacking energies and the loss of close packing for cytosine versus guanine.

G10T complex (K_rel = 12). Similar to G10C, the thymidine base replaces the cognate guanine with relatively little change (Figure 5F). In the cognate complex, the guanine N7 group hydrogen bonds to βK145; this hydrogen bond is maintained in the non-cognate complex in which the O2 hydrogen bonds to βK145. The indirect water-mediated interaction seen in the cognate complex is also maintained.
Fig. 5. Complexes with G10 nucleotide substitutions. Hydrogen bonds are shown as solid yellow sticks. Water molecules are shown as small turquoise spheres. The backbone amide of BA110, mentioned in the text, has been omitted in all images for clarity. (A) Detail of the cognate protein–ssDNA interactions at the G10 position. G10, in the syn conformation, is bound between αY239 and βR140, with two hydrogen bonds shown between the base and the protein. (B) G10l: note the absence of a hydrogen bond with the backbone carbonyl of βS102. (C) G10X: the protein remains in a conformation similar to cognate. An exception is αY239, which is rotated into a preferred rotamer conformation. A water molecule replaces the G10 guanine and interacts with βK145. (D) G10A: no DNA–protein hydrogen bonds are observed. The adenine is in the anti conformation, flipped 180° relative to the guanine. (E) G10C: two fewer hydrogen bonds, relative to the cognate complex, result in a similar protein conformation, maintaining all stacking interactions. (F) G10T: One hydrogen bond is similar to the cognate complex (between the base and βK145), while a novel hydrogen bond is made between the thymine O2 and αT240.

Fig. 6. G11 nucleotide substitutions. (A) View of the cognate protein–ssDNA interactions at the G11 position. G11 is bound in a deep cleft between two domains of the α subunit, surrounded by hydrophobic residues. A direct hydrogen bond is made between the G11 base and αK261. The base and αH14 interact via a bound water molecule. (B) G11C and (C) G11T: a water molecule is involved in a three-way indirect interaction between the bases, αK261 and αH14. The pyrimidines are tilted 15° relative to the cognate guanine.

The G10T substitution results in two noteworthy differences: (i) the loss of a single direct hydrogen bond at the protein–DNA interface (guanine N2 to the βS102 backbone carbonyl), and (ii) the introduction of a novel interaction between the DNA and protein (thymine O2 to the αT240 hydroxyl group). The thymidine methyl group is exposed to solvent.

Intriguingly, the G10T substitution results in a slightly smaller yet statistically significant energetic loss (ΔΔG_rel = 6.2 kJ/mol; P < 0.05) than either the G10A or the G10C substitution. It is perhaps noteworthy that the G10T substitution maintains two direct hydrogen bonds with the protein, in contrast with the G10A and G10C complexes which both have fewer hydrogen bonds.

Interactions with G11
The G11 was chosen because it is bound deep within a markedly hydrophobic pocket between the two OB folds of the α-subunit in an anti conformation (Figure 6A), features that may contribute to specificity for a guanine base. The two faces of the G11 base pack against several aliphatic and aromatic protein side chains. A direct hydrogen bond is made between the backbone carbonyl of αK261 and the exocyclic amino group of G11, which in turn participates in a water-mediated interaction with a ring nitrogen of αH14.

G11C complex (K_rel = 81). As with the G10 substitutions, the G11C substitution is accommodated by the ternary
complex with only minor structural changes. The protein side-chain conformations are indistinguishable from those of the cognate complex. As shown in Figure 6B, the plane of the cytosine base has rotated about the glycosidic bond ~15° with respect to the plane of the cognate guanine, resulting in a more perpendicular orientation of the base with respect to the attached ribose ring.

The cytosine O2 makes a water-mediated contact with a ring nitrogen of αH114. This water molecule is well ordered in the G11C crystal structure and interacts with the backbone carbonyl of αK261 to form a three-way interaction at the protein–DNA interface. This water-mediated interaction has replaced a direct hydrogen bond that previously existed between the exocyclic amino group of the cognate G11 guanine and the backbone carbonyl of αK261. Therefore, a water molecule appears to help mediate this site for interaction with a non-cognate cytidine nucleotide. The G11C substitution results in a larger reduction in binding free energy (∆G<sub>rel</sub> = 11 kJ/mol) than the equivalent substitution at G10, perhaps due to the greater hydrophobic character of the G11 binding pocket.

**G11T complex (K<sub>eq</sub> = 44).** The ternary complex also accommodates the G11T substitution with minor structural changes (Figure 6C). Virtually identical to the G11C complex, the thymine base has rotated about the glycosidic bond by ~15°, while the thymine O2 is involved in a three-way water-mediated interaction at the protein–DNA interface.

G11T also results in a somewhat larger reduction in binding free energy (∆G<sub>rel</sub> = 9.4 kJ/mol) than the corresponding substitution at the G10 position, which again may reflect the greater hydrophobic nature of the G11 binding site.

**Discussion**

We have explored the molecular basis of cognate telomeric ssDNA recognition via structural and thermodynamic comparisons of 10 different non-cognate OnTEBP complexes, two of which include ssDNAs that correspond to sequences found during the process of telomeric polymerization. This analysis has revealed an unanticipated plasticity in accommodating non-cognate sequences, including conformational rearrangements involving shifts in the ssDNA register of up to nine nucleotides, a phenomenon that we have termed nucleotide shuffling. The conformational perturbations that result from individual base changes in the cognate sequence are accompanied by significant, yet relatively modest, reductions in binding affinity. Without the structural data, these energetic differences might have been misinterpreted as reflecting loss of binding interactions in an otherwise unchanged ssDNA register. Instead, we show that the complex often finds a more energetically favorable solution for dealing with the altered nucleotide sequence, often involving contacts far removed from the mutated site.

Relative to dsDNA, single-stranded nucleic acids are very flexible and the ssDNA phosphodiester backbone can explore a large conformational space. Via various conformational rearrangements that exploit this flexibility, a non-cognate ssDNA can adjust its sequence register to preserve cognate-like interactions and maximize the energetics of favorable contacts. By using the different techniques discussed below, the strict requirement for a specific linear sequence of bases can be eliminated, often without substantial energetic penalty.

In some structures, such as the G3X, Tel+T and Tel-3 complexes, a specific base has been substituted, deleted or modified, yet all cognate interactions at the corresponding binding site are maintained because a neighboring base is able to replace the non-cognate base. In these three non-cognate structures, a large fraction of the energetic effects of the ssDNA base substitutions arise from structural perturbations that are distant from the cognate mutated site. Therefore structural data is essential for correctly interpreting the physical basis of our thermodynamic perturbation studies. On the other hand, the magnitude of structural change observed in the ssDNA and protein does not necessarily correlate with a similar change in binding affinity, a fact that underscores the importance of binding studies in dissecting the energetic importance of interactions observed in high-resolution structures.

Local sequence context is critical for determining the effects of a particular base substitution. For example, a register shift and base expulsion is conceivable in the G10X complex, as is seen in the G3X complex. However, the interaction energy contributed by adjacent bases (e.g. T8, G9, G11 and G12) appears to outweigh the effects of removing G10 and thereby prevents a shift in G10X DNA register. In complexes where the register does not shift, base stacking is always maintained (both with other ssDNA bases and with aromatic protein side chains), likely reflecting the energetic importance of stacking interactions in protein–ssDNA complexes.

Water molecules and hydrogen bonds are used creatively in these complexes to accommodate various base changes. Water-mediated interactions can substitute for direct protein–ssDNA hydrogen-bond interactions (e.g. G11C and G11T complexes). In the abasic G10X complex, water fills the cavity left by removal of the base. Novel and seemingly specific hydrogen-bond interactions can be formed in non-cognate complexes (such as seen with G10T). In other cases, cognate-like hydrogen-bond interactions are preserved despite the substitution of a non-cognate base.

Nucleotide shuffling has important consequences for protein recognition. In double-stranded DNA–protein complexes, single base-pair changes from the cognate sequence frequently result in very large reductions in binding affinity. Because one base pair can account for nearly the entire binding energy in dsDNA systems, the energetic contribution of single nucleotides to the overall specific binding affinity is not additive but highly cooperative (Lesser et al., 1990; Takeda et al., 1992; Englert et al., 1997; Frank et al., 1997). In contrast, we observe that single base changes in the cognate ssDNA result in modest decreases in binding affinity that are consistent with additivity. Detailed binding studies with OnTEBP have confirmed this observation (Gray, 1992; Classen et al., 2003), and similar results have been seen in several other ssDNA– and ssRNA–protein complexes, including Cdc13, PTB, SF1, U2AF<sup>65</sup> and Sex-lethal (Berglund et al., 1997; Perez et al., 1997; Singh et al., 2000; Anderson et al., 2003). Therefore, unlike what is
commonly observed with sequence-specific dsDNA–protein complexes, low cooperativity between cognate nucleotides may be a general characteristic of sequence-specific recognition of single-stranded nucleic acids. In light of our crystal structures with non-cognate ssDNAs, this feature can be straightforwardly interpreted as the consequence of multiple near-isenergetic conformations available to many similar ssDNA sequences.

The generality of this principle is suggested by three protein–RNA complexes, two involving hairpins with short RNA stretches comparable to a single-stranded nucleic acid. Alternate nucleic acid conformations, albeit less extensive than those described here, are also seen with an unnatural nucleotide in the RNA tetraloop of the MS2 coat binding protein complex (Grath et al., 2000) and in the Nova KH domain complexed with a stem loop SELEX RNA involving nucleotides, outside the core UCAC recognition motif, with no or few protein contacts (Lewis et al., 2000). A subtle conformational rearrangement has also been observed in the ssRNA ‘spacer’ region of the TRAP–ssRNA complex (involving an inserted base lacking direct protein contacts) (Hopcroft et al., 2002).

These observations emphasize the fundamental differences between protein recognition of single-stranded versus double-stranded nucleic acids. Cognate complexes with double-stranded nucleic acids have a high degree of complementarity with few waters at the protein/nucleic acid interface (Garnier and Rau, 1995; Sidorova and Rau, 1996) and involve local folding of the protein (Spolar and Record, 1994). Non-cognate complexes in these systems have quite different structures, such as poor interface complementarity with a large solvent content and an absence of local folding (Record et al., 1991; Engler et al., 1997; Oda et al., 1998). In contrast, both the cognate OnTEBP–ssDNA complex and the non-cognate complexes described here have extensive comparable shape complementarity at the protein–ssDNA interface with few water-mediated contacts. Similarly, there is no obvious pathway for the ssDNAs to enter the protein binding site if the OnTEBP protein surface is preformed. Thus both cognate and non-cognate OnTEBP complex formation appears to proceed by a cofolding pathway (Horvath et al., 1998), in which protein and nucleic acid fold together to form the final complex.

Telomeric end protection is a critical function of proteins that specifically bind to the telomeric single-stranded overhang, including OnTEBP, Schizosaccharomyces pombe Pot1, and Saccharomyces cerevisiae Cdc13 (Gottschling and Zakian, 1986; Nugent et al., 1996; Baumann and Cech, 2001). The combination of sequence, structural and functional similarity among these divergent eukaryotic proteins indicates that mechanisms of telomeric end protection are broadly conserved throughout evolution (Mitton-Fry et al., 2002; Theobald et al., 2003). Interestingly, two of the ssDNAs studied in this work, Tel+T and Tel-3, correspond to intermediates generated during extension of the telomere by the enzyme telomerase. Although it is unclear whether OnTEBP is present at the telomere in vivo when telomerase adds repeats, OnTEBP binding to the telomere inhibits telomerase activity in vitro, indicating a role for OnTEBP in telomere length regulation (Froelich-Ammon et al., 1998). Despite the limited in vitro tolerance of

OnTEBP for non-cognate sequences, the length and sequence identity of O.nova telomeric ssDNA is closely regulated in vivo (Klobucar et al., 1981). Taken together, these observations suggest that the widespread biological preference for specific telomeric repeats and guanosine-rich telomeres largely results from factors other than these telomeric end-binding proteins, with telomerase a likely candidate.

Nucleotide shuffling may also explain how a variety of other proteins recognize heterogeneous single-stranded sequences. For example, the yeast telomeric proteins Cdc13 and Pot1 bind irregular G-rich ssDNA repeats, and splicing factors such as PTB, U2AF and Sex-lethal recognize heterogeneous ssRNA sequences (Singh et al., 2000). Several known in vivo PTB binding sites (e.g. human α-tropomyosin IVS2 and IVS3 and rat β-tropomyosin IVS6) differ from the PTB consensus sequence by a single-base insertion. This puzzling feature could result from the ability of the pre-mRNA to shift register and flip an unfavorable base into solution. Similarly, much of the variability in U2AF binding sites can be readily explained by invoking nucleotide shuffling, especially the single-base differences observed in the conserved internal CCC tract.

**Materials and methods**

Nitrocellulose filter-binding assays were used to determine relative $K_D$ values for the 10 non-cognate ssDNA sequences (Table I). The ssDNAs used in these filter-binding experiments each had an additional 28 nucleotides (‘Onx’ = AAGACGACATTCCTCAGCCAGACATTI) at the 5’ end, included to pre-empt G-quartet side reactions and to facilitate efficient binding of free ssDNA to the bottom filter. Weak and non-specific interaction of the 5’ tail with the protein (RaghuRaman et al., 1989) was abrogated by addition of non-specific competitor ssDNA. Analogous binding experiments using an agarose gel system with the shorter ssDNAs, though more difficult, yielded qualitatively similar results. Final binding reactions contained 25 mM HEPES pH 7.5, 150 mM NaCl, 5% glycerol, 0.01% Tween-20, 2 mM dithiothreitol (DTT), 1 mM EDTA, 10 µM non-specific competitor ssDNA (TelRNDM) and 50–200 pM 5’-32P-labeled ssDNA. Reaction mixtures were allowed to equilibrate at room temperature for 24 h and were filtered using a two-membrane Schleicher & Schuell 96-well minifold dotblot apparatus with a nitrocellulose top layer and a Hybond-N+ (Amersham-Pharmacia) bottom layer. The β-subunit contains a 28 kDa core domain ([β28], used here, that is sufficient for ternary complex formation and that lacks G-quartet binding activity (Fang and Cech, 1993a, 1993b; Fang et al., 1993).

In this ternary system, the unitless $K_D$ is formally defined as

$$K_D = \frac{[\gamma_0][\alpha][\beta][\gamma_{28}[DNA]]}{[\gamma_0][\beta_{28}-DNA][\alpha][\beta_{28}-DNA]}$$

where activity coefficients $\chi$ are ~1.0/M at low concentration. Binding constants were fitted to

$$f = \frac{[\alpha][\beta][\gamma_{28}] - [\alpha][\beta][\gamma_{28}]}{[\alpha][\beta][\gamma_{28}] + K_D}$$

where $f$ is the fraction of DNA bound. Concentrations of $\alpha$ and $\beta_{28}$ were held equivalent, and thus at half-bound DNA, the $K_D$ equals the product of $[\alpha]$ and $[\beta_{28}]$. All binding experiments were performed in triplicate. For cognate OxyTel, $K_D = (2.8 \pm 0.3 \times 10^{-9})^2$, consistent with previously reported measurements (RaghuRaman et al., 1989; RaghuRaman and Cech, 1990; Gray, 1992; Fang et al., 1993).

For the $K_{cog}$ competition assay, a series of binding reactions were made in which the protein and radiolabeled cognate ssDNA were held at constant concentration and combined with increasing amounts of competitor ssDNA (TelRNDM, final concentrations 10 µM–1 mM). Labeled ssDNA was in slight (~10%) molar excess over protein (1–2 µM). Reactions were incubated at room temperature for 24–48 h before agarose gel electrophoresis. Results were fitted to the equation

$$(1 - f)/(f - 1) = [\text{TelRNDM}]_0/[K_{cog} + [\text{Tel ssDNA}]_0]$$
where $j$ is the fraction of DNA bound and $f_i$ is the fraction bound with no competitor ssDNA.

The α- and β-protein subunits were expressed in Escherichia coli and purified as previously described (Horvath et al., 1998). Ternary complex with modified ssDNA was crystalized as described previously (Horvath et al., 1998) with one major alteration based on the discovery that intact G-quartets were necessary for crystallization. Briefly, G-quartets of a 13-nucleotide ssDNA (2.86 mM G_{13}T_{13}G_{13}T) were preformed by annealing from 100°C to room temperature, followed by incubation in 100 mM NaCl for 3 weeks. Ternary complex was formed first (200 μM α-subunit with 1.1 to 1.4 molar equivalents of J28 and modified ssDNA), followed by addition of the preformed G-quartets (250 μM final). Ternary complex (~2 μL) was added to an equal volume of precipitant (30% ethylene glycol, 1–6% polyethylene glycol 4000, 40 mM MES pH 5.5–7.0, 0.002% NaN₃, 2 mM DTT) and allowed to equilibrate against a 1 ml volume of precipitant with an additional 50 mM NaCl. Crystal morphology of non-cognate complexes was indistinguishable from cognate complexes (space group P6₁22 with a = b = 94 A and c = 425 A).

Diffraction data were collected at the synchrotron beamline 14BM-C (Argonne APS) from cryoprotected crystals. Images were processed with HKL (Otwinowski and Minor, 1997). All structures were solved by molecular replacement (MR) using CNS (Brünger et al., 1998). Phases were generated from two complementary models derived from the cognate OssTEBP-Tel complex structures (1OTC and 1JB7: an OssTEBP-Tel model containing the ssDNA and one omitting the ssDNA. Rigid body refinement followed by rounds of minimization, simulated annealing and manual refinement reduced $R_{	ext{free}}$ from >0.4 to <0.3. Differences between cognate and non-cognate complexes were clearly revealed in F_0-C_F maps. As a precaution to eliminate bias, the ssDNAs in the OssTEBP-Tel and OssTEBP-Tel-3 models were built independently of the cognate model, using phases calculated from the MR model omitting the Tel ssDNA. No crystal contacts involving the ssDNAs are observed, though there is a close approach to a symmetry-related molecule near G1. Comparison with independent crystal structures of α:ssDNA complexes (1K8G and 1K1T) indicates that at least two adjacent β cognate ssDNA binding sites are unobstructed in the ternary complex crystals, suggesting that it is unlikely this symmetry-related molecule would influence binding of these non-cognate ssDNAs.

Acknowledgements

We thank Thomas Cech and Deborah Wutzke for critical review of the manuscript and Catherine Theobald for contribution to figure design. The American Cancer Society (RPG-93–011–04–NP) and the NIH (1R01CA81109) provided support.

References


Horvath,M.P. and Schultz,S.C. (2001) DNA G-quartets in a 1.86 A


Spolas,R.S. and Record,M.T., Jr. (1994) Coupling of local folding to site-specific binding of proteins to DNA. Science, 263, 777–784.


Received January 27, 2003; revised June 27, 2003; accepted July 2, 2003