1999

Novel G-DNA binding proteins from the ciliate Tetrahymena thermophila: purification, characterization, cloning and functional analyses

Quan Lu
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd
Part of the Cell Biology Commons, and the Molecular Biology Commons

Recommended Citation
Lu, Quan, "Novel G-DNA binding proteins from the ciliate Tetrahymena thermophila: purification, characterization, cloning and functional analyses" (1999). Retrospective Theses and Dissertations. 12469.
https://lib.dr.iastate.edu/rtd/12469

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA

UMI®
800-521-0600
Novel G-DNA binding proteins from the ciliate *Tetrahymena thermophila*:
Purification, characterization, cloning and functional analyses

by

Quan Lu

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Molecular, Cellular and Developmental Biology

Major Professor: Eric Henderson

Iowa State University

Ames, Iowa

1999
Graduate College
Iowa State University

This is to certify that Doctoral dissertation of

Quan Lu

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program

Signature was redacted for privacy.

For the Graduate College
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iii</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>CHAPTER 1: PURIFICATION, CHARACTERIZATION AND MOLECULAR</td>
<td>11</td>
</tr>
<tr>
<td>CLONING OF TGP1, A NOVEL G-DNA BINDING PROTEIN FROM</td>
<td></td>
</tr>
<tr>
<td>TETRAHYMENA THERMOPHILA</td>
<td></td>
</tr>
<tr>
<td>CHAPTER 2: TWO NOVEL TETRAHYMENA G-DNA BINDING PROTEINS,</td>
<td>34</td>
</tr>
<tr>
<td>TGP1 AND TGP3, SHARE REPETITIVE AND HOMOLOGOUS DOMAINS</td>
<td></td>
</tr>
<tr>
<td>CHAPTER 3: FUNCTIONAL ANALYSES OF TWO NOVEL TETRAHYMENA</td>
<td>47</td>
</tr>
<tr>
<td>G-DNA BINDING PROTEINS: A ROLE IN MICRONUCLEAR DIVISION</td>
<td></td>
</tr>
<tr>
<td>GENERAL CONCLUSIONS</td>
<td>70</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>72</td>
</tr>
</tbody>
</table>
ABSTRACT

G-DNA is a family of novel four-stranded DNA structures characterized by motifs called G-quartets. Evidence is growing to suggest that G-DNA exists and plays biological roles in vivo. In order to further elucidate the functions of G-DNA, we have studied proteins that specifically bind to the DNA structure.

Two G-DNA binding proteins, TGPI and TGP3, were purified from the ciliate **Tetrahymena thermophila**. Based on the peptide sequences obtained from direct internal peptide sequencing, the cDNAs coding for the genes were cloned. Deduced protein sequences showed that TGPI and TGP3 are novel proteins but share significant homology with each other. Furthermore, the two proteins contain an intriguing sequence pattern with two repetitive and homologous motifs flanking an extensively hydrophilic region. We suggest that this shared sequence pattern may represent a novel G-DNA binding motif.

To address the biological functions of these two novel G-DNA binding proteins, we have employed a newly developed gene knock out technique and generated **Tetrahymena** strains with each of the two genes completely disrupted in the macronucleus. Both knock-out strains (TGPIKO and TGP3KO) grow normally, suggesting that neither of the genes is essential for cell growth and survival. Nevertheless, detailed nuclear staining analysis revealed micronuclear aberrance characterized by higher occurrence of multiple micronuclei in both knock-out cells, suggesting a faulty control of the micronuclear division. More interestingly, TGPIKO cells showed an increased TGP3 activity, implying that these two proteins may share some aspects of biological functions. In addition to the gene knock-out experiments, we also did nuclear fractionation experiment that demonstrated that both TGPI and TGP3 localize mainly in the nuclei. Based on these data, we propose a model in which TGP proteins coordinate to function in micronuclear division through binding to the G-DNA structure formed between telomeres of two sister chromatids.
INTRODUCTION

Dissertation organization

The dissertation contains three major parts: an introduction, the main body and a general conclusion. In the introduction, I give an overview of an unusual four-stranded DNA structure known as the G-DNA and its binding proteins with a focus on their biological implications. I also briefly introduce the model system I used. The main body of this dissertation is composed of three papers (Chapter 1-3). The first paper, which has been published in *Nucleic Acids Research*, is about the purification, characterization and cloning of a novel G-DNA binding protein called TGP1. The second paper is about the cloning of an additional G-DNA binding protein TGP3, which turns out to share significant homology with TGP1. This paper has been submitted for publication. The third chapter is a "manuscript in preparation" concerning the *in vivo* roles of these two novel proteins. I am the primary author on each of these papers and have performed most of the experiments described. In the general conclusion, I summarize results of my research and address topics worthy of further investigation. References cited in the introduction and conclusion are listed together after the conclusion.

Unusual DNA structures

DNA is a flexible molecule capable of adopting different structures or conformations. The primary DNA conformation in a cell is the B-form DNA, which is a right-handed double-helix with 10.4 bases per turn. At least *in vitro*, double-stranded DNA can adopt alternative conformations such as the dehydrated A-DNA, left-handed Z-DNA and parallel-stranded duplex DNA (1). Recently, it was shown that B-DNA can even be stretched and twisted to form a dramatically different structure called P-form DNA, which has the phosphate backbone on the inside of the duplex (2). In addition to these double helix conformations, multiple-stranded DNA structures (triplex and tetraplex) have been shown to exist. Different from duplex DNA, these multiple-stranded DNA structures are usually stabilized by non-Watson-Crick base pairings. For example, a triplex DNA is formed when pyrimidine or purine bases form Hoogsteen base pairs with purines of the Watson-Crick base pairs (3). Tetraplexes are mediated by either protonated C-C+ base pairs (C-tetraplex, (4,5)) or guanine-guanine base pairings (G-tetraplex, (6)). The G-tetraplex is the focus of my thesis research.

G-quartet motif and G-DNA

The major structural feature of the G-tetraplex is a novel motif known as the G-quartet, which was first proposed more than three decades ago by Gellert to explain the sticky gel formation of guanosine and its derivatives (6). The G-quartet motif is a planer, tetrameric arrangement of four guanines, each of which forming G:G hydrogen bonds with its two
neighbors (Fig. 1; 6,7). Since their large planar surfaces would result in strong van der Waals attractions, several G-quartets can stack upon each other to form a four-stranded helical structure (G-tetraplex). G-tetraplex structures can be stabilized by the presence of certain monovalent cation molecules such as potassium and sodium.

G-tetraplexes usually arise from guanine-rich sequences containing clusters of guanines. Depending on the primary DNA sequence (and other factors such as ionic environment and DNA concentration), different forms of G-tetraplex can be adopted constituting a closely related family ((8); Fig. 2). For example, G-tetraplex can be formed by clustering of four individual DNA strands (tetrameric), annealing of two DNA hairpins (dimeric), or foldback of a single DNA strand (monomeric). The DNA strand orientation in these structures can be parallel or anti-parallel. For the purpose of clarity, we refer to all these G-tetraplex structures collectively as G-DNA.

Telomeres can form G-DNA in vitro

Interest to study G-DNA has surged since it was discovered that telomeric DNA can from G-DNA structures in vitro (7,9). Telomeres are the terminal nucleoprotein structure of eukaryotic chromosomes (10,11). They protect chromosomes from degradation and end-to-end fusion, and thus are essential for chromosome stability and integrity. Telomeres may also play a role in chromosome organization within the nucleus as suggested by telomere clustering at the nuclear periphery (12). In most organisms, telomeres consist of simple short DNA repeats (Fig. 3A). These short repeats are usually G/C rich with the 5'->3' strand (toward the chromosome end) always guanine-rich and extending beyond the double-stranded part to form a single-stranded 3'-overhang ((7), Fig. 3B). These features of telomere sequences are generally conserved among different organisms and are thought to be related to telomere function.

It was first observed that, under appropriate ionic conditions, naked (devoid of proteins) telomeres of *Oxytricha* macronuclear DNA can cohere together at high concentrations (13,14), suggesting that telomere DNA may be capable of adopting unusual structures. This suggestion was confirmed by two pioneering studies of telomere DNA structures formed by synthetic telomeric oligonucleotides (oligo). In 1987, Henderson and his colleagues (15) found that several guanine-rich telomeric oligos including the *Tetrahymena* telomeric oligo d(T4G4)4 can form very compact structures that migrate abnormally on nondenaturing gels. Subsequent NMR analysis suggested that G:G base pairs exist in the unusual compact structures. Two years later, Williamson et al. (16) confirmed the involvement of G:G base pairing in the folded structure of *Oxytricha* telomeric oligo d(T4G4)4. They observed an increased electrophoretic mobility of the oligo in nondenaturing gels. In addition, they showed that the formation of folded structure in the oligo results in the protection of guanines from methylation by dimethyl sulfate (DMS), suggesting that
Figure 1. Schematic diagram of G-quartet motif and G-DNA structure. G-DNA containing stretches of G-quartet motifs is stabilized by base stacking and cation coordination. Adopted from Henderson (7).

Figure 2. Diversity of G-DNA structures. The arrows indicate 5' to 3' strand orientation.
the guanines are involved in hydrogen pairing. According to these data, G-DNA was proposed to be formed by these telomeric oligos.

Numerous investigations following these early studies further confirmed the G-DNA formation by telomeric DNA (3,7,9). By using oligos of different sequences and lengths, these studies also demonstrated that telomeric oligos can adopt a variety of forms of G-DNA. For example, NMR study by Smith and Feigon revealed formation of intramolecular foldback G-DNA in the oligo (G,T)3G4 (17). A shorter form of this oligo d(G,T,G4) was found to form a dimeric (bimolecular hairpin) G-DNA (17), which was confirmed by an independent X-ray crystal study ((18), Fig. 3C). Telomeric oligos such as d(TG,T) can form tetrameric G-DNA (19). Intriguingly, a short telomeric oligo d(G,T,G4) can also form large superstructures containing G-quartets (20). These superstructures known as the G-wire are about 1 μm long and can be visualized by the atomic force microscopy (21).

Non-telomeric sequences that form G-DNA

Like telomeres, many other chromosomal domains contain sequences with blocks of guanines. These domains include the immunoglobulin switch region (22), mouse hypervariable microsatellite site (23), the fragile X site (24) and several gene promoters (25-29). The immunoglobulin switch region contains stretches of repetitive DNA with conserved motifs such as GGGGT and GAGCT (22). Synthesized oligos corresponding to these sequences spontaneously form complexes of very low electrophoretic mobility (30). Dimethyl sulfate (DMS) protection analysis indicated formation of G:G base pairing in the complexes. It was proposed that formation of such G-DNA structure will bring different constant regions together and thus may play a role in switch recombination (30). The mouse hypervariable microsatellite consists of a tandem array of d(CAGGG)n (23). In vitro, this sequence can form a hairpin as well as tetraplexes characterized by G-quartets (31). Formation of these structures may affect DNA replication and recombination, and thus contribute to the instability property of these hypervariable regions. The fragile X locus contains triple nucleotides (CGG)n repeats and the expansion of these CGG repeats has been linked to the X fragile syndrome. These repeats are capable of forming G-DNA in vitro (24). Finally, guanine-rich sequences are also found in promoters of several genes including chicken β-globin gene (25,27), cysteine proteinase genes (26), retinoblastoma susceptibility genes (28) and c-myc oncogene (29).

G-DNA binding proteins

Generally, there are two approaches to study G-DNA in vivo. First, G-DNA may be studied directly with the aid of G-DNA specific reagents such as G-DNA antibodies (32,33) or specific fluorescent dyes (34). However, the usefulness of these reagents in detecting G-DNA in
**Figure 3.** (A) Telomeric DNA sequences from most species are G-rich. (B) *Tetrahymena* telomeric terminus has a 3'-overhang. This overhang feature is conserved among telomeres from many organisms. (C) Crystal structure of G-DNA formed by *Oxytricha* telomeric DNA (18).

**Figure 4.** Possible G-DNA structures at telomeres. These different G-DNA structures may serve different functional roles. See text for details.
vivo still remain to be seen. An alternative way to study G-DNA is to investigate protein factors interacting with the structure.

Several proteins with G-DNA binding activity have been identified. Some of these G-DNA binding proteins are previously known proteins. For example, the first protein found to exhibit G-DNA binding activity was MyoD, the transcription factor known to function in the initiation of myogenesis (35). It was shown that recombinant MyoD specifically binds to helical structure formed by stacks of guanine residues in G-quartet motif (36). Macrophage scavenger receptors are trimeric integral membrane glycoproteins that have been implicated in the deposition of lipoprotein cholesterol in artery walls. The receptors are able to bind to polynucleotides in G-DNA structure (37). Such G-DNA polynucleotide binding inhibits the activity of the receptors (37). Dihydrolipoamide dehydrogenase from several species including Tetrahymena were found to bind to G-DNA (38). The yeast cytosolic glutamyl-tRNA synthase (G4p2) was also found to contain G-DNA binding activity (39). Recently, LR1, a B-cell specific protein that can bind to the immunoglobulin switch region, was found to be able to bind to G-DNA (40). In addition, novel G-DNA-binding proteins have been identified. These novel proteins include a hepatocyte chromatin protein termed QUAD (41), the yeast G4p1 protein (42), and two Tetrahymena proteins TGPl and TGP3 (43,44).

Two proteins were identified which not only bind to G-DNA but also cleave the G-DNA substrate once bound. One of the proteins is the yeast nuclease KEM1. Liu et al. first identified a G-DNA specific nuclease activity that binds to G-DNA structure regardless of the primary sequence and then cleaves the single-stranded region 5' to the G-quartet motif (45). This nuclease activity was later purified and shown to be the product of the yeast KEM1 (also known as SEPl, DST2, XRN1 and RAR5) gene (46), a gene essential for meiotic recombination (47). The other G-DNA-cutting protein is the eukaryotic topoisomerase II. In vitro assay showed that topoisomerase cannot cut single-stranded molecule but can cleave the same DNA molecule in G-DNA conformation (48).

Proteins catalyzing G-DNA formation or unwinding

Formation of G-DNA structure from single-stranded DNA is a thermodynamic reaction. Two proteins have been found to promote the G-DNA formation under physiological conditions in vitro. One of the proteins is the Oxytricha telomere-end binding protein (TEBP) β-subunit, which enhances the rate of G-DNA formation by over 100 fold at a DNA concentration of 20 nM (49). The TEBP β-subunit acts as a molecular chaperone since it promotes G-DNA formation without remaining stably bound to the DNA substrate. Protein deletion analysis indicates that the C-terminal region of the TEBP β-subunit is responsible for the G-DNA promoting activity (50). Like the TEBP β-subunit, the yeast RAP1 protein can also facilitate the formation of G-DNA (51). RAP1 is a multifunctional protein essential for the maintenance of yeast telomeres (52). It
was found that RAP1 not only bind to double-stranded DNA, but also interacts with G-rich strand containing G-quartets. Circular dichroism spectroscopy analysis suggested that RAP1 promotes formation of G-DNA by bringing together the G-rich strands and thus increasing the local concentration of single-stranded DNA substrate (51).

In contrast to the G-DNA-promoting proteins, three proteins have been identified to be able to do the opposite: they unwind G-DNA into single-stranded DNA. One of such G-DNA-unwinding protein was found in human placental tissue (53). Biochemical characterization showed that this protein is not a helicase or topoisomerase and does not unwind Hoogsteen-bonded triplex DNA (53). Different from this protein, two other G-DNA unwinding proteins are previously identified helicases: the SV40 large T-antigen helicase (54) and the Bloom's Syndrome helicase (55). The G-DNA unwinding activity of both helicases requires ATP and a short single-stranded tail 3' to the G-quartet region (54,55). Interestingly, both helicases prefer G-DNA to double-stranded DNA substrate. Identification of these G-DNA unwinding activities has important implications. G-DNA are usually very stable once formed, leading to the suggestion that G-DNA may not exist in vivo since its high stability may impede with normal biological processes. However, identification of G-DNA-unwinding proteins suggests that the cell may use these proteins to resolve G-DNA structures in vivo.

**Putative biological roles of G-DNA**

Identification of biological sequences capable of forming G-DNA and protein factors interacting with G-DNA supports the hypothesis that G-DNA structures exist and play roles in vivo. It is proposed that G-DNA may function in many diverse biological processes including telomere function, recombination and gene regulation.

**G-DNA at telomeres.** As we discussed earlier, telomeric DNA can form G-DNA in vivo under physiological conditions, and two telomere-binding proteins (the *Oxytricha* TEBP β-subunit and the yeast RAP1 protein) can even act as molecular chaperones to promote the formation of G-DNA, strongly suggesting that G-DNA may present and play roles at telomeres. There are several ways for G-DNA to form and function at the telomeres (Figure 4). The first and maybe most possible way is that the 3’-G-overhangs of two telomeres associate together to form a dimeric G-DNA. Such a G-DNA structure may be responsible for telomere-telomere association, which has been shown to play a role in chromosome segregation during mitosis (56,57). The second way of G-DNA formation could be that the telomere G-strand folds back to form intramolecular (monomeric) G-DNA structures. Since G-DNA is usually more resistance to nuclease digestion than double-stranded DNA, these G-DNA nodes may serve a structural role to protect the chromosome from degradation. Finally, G-DNA can also be formed by annealing of four G-overhangs (or G-strands) of telomeres. This tetrameric G-DNA may serve a role in homologous chromosome alignment during meiosis (58). All forms of G-DNA formed at
telomeres may be recognized by certain specific G-DNA binding proteins in vivo. The resulting G-DNA/protein complexes may serve several potential roles. For example, the complexes may regulate the telomere length by limiting the access of telomeres to telomerase (59). The complexes may also serve as anchors to link telomeres to specific nuclear structures such as the nuclear periphery (12), thus helping organize chromosomes in the nucleus.

**G-DNA in recombination.** The most convincing evidence for G-DNA's role in recombination came from studies of the yeast KEM1 protein. Homologous KEM1 deletion results in blockage of meiosis at the 4N stage (47). FACS analysis revealed that the mutants underwent premeiotic DNA synthesis but arrested before spore wall formation, suggesting that a role for KEM1 in meiotic recombination. The KEM1 protein was later purified and found to be a G-DNA dependent nuclease (46) and contain strand-exchange activity (60). A model was proposed to account for these activities of KEM1 (46) (Fig. 5). In this model, homologous chromosomes are tied together through a G-DNA structure which can be recognized by a putative KEM1-containing recombinase machinery. The nuclease activity of KEM1 cleaves the DNA strands to be recombined, and the strand-exchange activity of the protein can then transfer sequences between the homologous chromosomes. This model may be a general mechanism responsible for other G-rich DNA mediated recombination such as the immunoglobulin switch recombination.

**Other roles of G-DNA.** G-DNA may play a role in gene regulation. As we mentioned earlier, several gene promoters have been found to be G-rich sequences, and the transcription factor MyoD was identified to be able to bind to G-DNA structure. A possible scenario for G-DNA to function in gene regulation would be that G-rich promoter region forms a G-DNA structure which can be recognized and bound by certain transcription factor. The transcription factor can thus control the expression of genes downstream of the promoter. As more sequences capable of forming G-DNA and proteins interacting with G-DNA are identified, new roles for G-DNA will likely emerge.

The ciliated protozoan *Tetrahymena thermophila*

*Tetrahymena thermophila* is a eukaryotic single-celled organism. Like other ciliates, it has two phases (vegetative and conjugation) in its life cycle, and contains two structurally and functionally differentiated nuclei: a germ-line micronucleus and a somatic macronucleus (61). The micronucleus is diploid (contains five pairs of chromosomes) and behaves much like nuclei found in other organisms, except that it is transcriptionally quiescent. In contrast, the macronucleus has a fragmented and greatly amplified genome which is transcriptionally active (synthesizing all the proteins in the cell). In addition, the macronucleus divides amitotically distributing chromosomes randomly into the daughter cells, resulting in a phenomenon known as
Figure 5. A model of G-rich DNA-mediated recombination. Homologous chromosomes are brought together by a G-DNA structure, which can be recognized and bound by a putative recombinase machinery. The recombinase could include proteins such as the KEM1 nuclease, which can carry out the cleavage and strand-exchange. This model is adopted from Liu et al. (46).
the phenotypic assortment in which a heterozygous macronuclear genome tends to become homogeneity after many generations of vegetative growth.

There are several advantages to use Tetrahymena as a model system (61). First, its unique features (e.g., nuclear dualism and unusual macronuclear genome organization) make it a good system to study processes such as genome rearrangement, gene excision, rDNA amplification and telomere function (61). Second, Tetrahymena cells are easy to culture and have a short generation time of about 2.5 hours. Third, development of DNA transformation techniques makes genetic manipulation possible in this system (62-64). Finally, gene knockout technique has been developed in Tetrahymena (65,66). This makes it possible to directly study to the function of a particular gene. Despite these good features, Tetrahymena system has some drawbacks. First, the genetic background in Tetrahymena is not as clear as that in the yeast, even though a Tetrahymena genome project has been initiated (Ed Orias, University of California, Santa Barbara). Thus, genetic analysis can not be easily performed in Tetrahymena. Second, Tetrahymena uses an unusual genetic codon system: TGA is the only stop codon in the organism, the other two "stop codons" (TAA, TAG) specify glutamine. This unusual codon usage makes it problematic to express Tetrahymena genes in other systems such as E. coli.

**Novel Tetrahymena G-DNA binding proteins**

Tetrahymena provides a good source of G-rich DNA since its macronuclear genome contains large amount of telomeres (T_G repeats) (67). These G-rich telomere DNA may adopt G-DNA structures in vivo. Our lab has identified three G-DNA binding proteins (TGP1, TGP2, and TGP3) from this organism (38,43,44,68). All the three proteins have been partially purified and biochemically characterized. In addition, cDNAs coding for two of the proteins (TGP1 and TGP3) have been cloned. While TGP2 was found to contain dihydrolipoamide dehydrogenase activity (38), TGP1 and TGP3 were shown to be novel proteins with no significant homologs in available protein databases (43,44). Nevertheless, TGP1 and TGP3 are homologous to each other, and share an sequence pattern which contains two repetitive and homologous motifs flanking an extensively hydrophilic and basic region (43,44). This intriguing sequence pattern may constitute a novel putative G-DNA binding domain. To investigate the biological functions of these two novel proteins, TGP1 (or TGP3) gene was completely disrupted in the somatic nucleus. Both gene knockout strains (TGP1KO and TGP3KO) grow at near normal rate suggesting that neither of the genes is essential for cell growth and survival. However, nuclear staining analysis revealed that, in both KO cells the percentage of cells containing multiple (>2) micronuclei is much higher than that in the wild-type cells, suggesting a faulty control of the micronuclear division in the KO cells. A model involved in TGP protein binding to the G-DNA formed between telomeres of sister chromatids was proposed to account for these data.
CHAPTER 1: PURIFICATION, CHARACTERIZATION, AND MOLECULAR CLONING
OF TGPI, A NOVEL G-DNA BINDING PROTEIN FROM TETRAHYMENA
THERMOPHILA


Quan Lu, Ted Schierer, Sang-Gu Kang and Eric Henderson
Department of Zoology and Genetics
Molecular, Cellular and Developmental Biology Program
Iowa State University
Ames, Iowa 50011, USA

ABSTRACT

G-DNA, a polymorphic family of four-stranded DNA structures, has been proposed to
play roles in a variety of biological processes including telomere function, meiotic recombination,
and gene regulation. Here we report the purification and cloning of TGPI, a G-DNA specific
binding protein from Tetrahymena thermophila. TGPI was purified by three column
chromatographies, including a G-DNA affinity column. Two major proteins (~80 kDa and ~40
kDa) were present in the most highly purified column fraction. Renaturation experiments showed
that the ~80 kDa protein contains TGPI activity. Biochemical characterization showed that TGPI
is a G-DNA specific binding protein with a preference for parallel G-DNAs. The TGPI/DNA
complex has a dissociation constant (Kd) of about 2.2 x 10^-4 M and TGPI can form supershift in
gel mobility shift assays. The cDNA coding TGPI was cloned and sequenced based upon an
internal peptide sequence obtained from the ~80 kDa protein. Sequence analyses showed that
TGPI is a basic protein with a pi of 10.58, and contains two extensively hydrophilic and basic
domains. Homology searches revealed that TGPI is a novel protein sharing weak similarities with
a number of proteins.

INTRODUCTION

G-quartets are unusual nucleic acid structures first described nearly three decades ago (1). In a G-quartet, four guanine molecules lie in a plane with each guanine forming G:G (instead of Watson-Crick A:T or G:C) hydrogen bonds with two adjacent guanines (1-3). DNA containing such structures is known as quadruplex, tetruplex or G-DNA (3). Many guanine-rich sequences of biological significance have been found to be capable of forming G-quartet structures in vitro under physiological conditions. Such G-rich sequences include most telomeres (4-6).
immunoglobulin switch regions (7), a few gene promoters (8-10), fragile X repeats (11), and the
dimerization domain in the human immunodeficiency virus (HIV) genome (12,13).

The ability to form G-DNA by these important sequences implies that G-DNA may be
biologically relevant. However, direct evidence for the existence of G-DNA in vivo is lacking. An
alternative approach to study G-DNA is to identify and investigate proteins that interact with this
structure. A number of proteins have been found to be able to bind to G-DNA. These proteins
include a hepatocyte chromatin protein QUAD (14), transcription factor MyoD (15), chick
topoisonerase II (16), two yeast proteins G4p1 and G4p2 (17,18), and the yeast KEM1 protein
(19). In contrast to all other G-DNA binding proteins, KEM1 not only binds to G-DNA, but also
acts as a nuclease to cut the single-stranded DNA 5' to the G-quartet domain (19,20). The G-
DNA specific nuclease activity of KEM1, together with the studies in which meiotic
recombination was found to be severely affected in KEM1 deletion mutants (21,22), suggests that
KEM1 and the G-DNA structure play roles in meiotic pairing (20). In addition to G-DNA
binding proteins, two proteins have been found to facilitate the formation of G-DNA structure:
the Oxytricha telomere binding protein β subunit, which acts as a molecular chaperone (23,24),
and the multifunctional yeast telomere protein RAP1 (25). Since both proteins are telomere-
related, their G-DNA promotion activities suggest that G-DNA may have a role in telomere
function.

The biological relevance of G-DNA was further suggested by several other studies
showing that G-DNA can inhibit the activities of a number of proteins. One of these G-DNA-
inhibited proteins is telomerase, the enzyme that synthesizes telomeres (26). It was found that
folding of telomere DNA (specifically the primer) into G-DNA inhibited the activity of
telomerase in vitro, suggesting that formation of G-DNA in the primer could down-regulate
telomere elongation in vivo (27). Another in vitro study showed that an antiparallel G-DNA
aptamer can inhibit thrombin activity (28). G-DNA was also shown to inhibit the activity of
macrophage scavenger receptors, the glycoproteins which may function in the deposition of
lipoprotein cholesterol (29). More interestingly, an oligonucleotide in G-DNA form was found to
be a potent inhibitor of HIV-1 integrase (30). These inhibitory activities of G-DNA suggest its
potential use as a pharmaceutical agent.

Possible biological functions for G-DNA are diverse, but remain to be definitively
demonstrated. Further studies on G-DNA and its related proteins are needed to elucidate the exact
role(s) of these structures in vivo. The ciliated protozoan, Tetrahymena thermophila, has proven
to be a useful system for identification of G-DNA binding proteins. This organism has two nuclei,
one of which contains a fragmented genome with thousands of telomeres (31). A relatively
abundant G-DNA binding activity has been identified in T. thermophila (32). The protein, TGP1
(Tetrahymena G-DNA binding protein 1), binds to an intermolecular G-DNA form of
d(TTGGGG)₄ under physiological conditions, and competition experiments showed that G-
DNAs competed strongly for TGPI binding, while non-G-DNA and G-RNA oligonucleotides did not (32). In this paper, we report the purification, molecular cloning and further biochemical characterization of TGPI. TGPI was shown to be an 83 kDa protein with binding activity specific for G-DNA. Complete cDNA sequence of TGPI was obtained and shown to encode a novel protein with two extensively hydrophilic and basic putative DNA binding domains and weak similarities to a number of other proteins. In light of these data, possible functions for TGPI are discussed.

**MATERIALS AND METHODS**

**Oligonucleotide synthesis, purification and labeling**

Oligonucleotides (oligos) (Table 1) were synthesized on an ABI 391 DNA synthesizer (ISU DNA facility), and purified as previously described (32). Briefly, oligos were separated on a 20% denaturing polyacrylamide gel (7M urea), then desired oligo products were excised and purified by C-18 chromatography (Waters). 5'-32P-labeling of oligos using T4 polynucleotide kinase was carried out according to a standard protocol (33). Radiolabeled oligos were purified by G-25 spin columns (5 prime→3 prime).

**Table 1. Oligonucleotides used in mobility shift and competition assays:**

Y: ACTGTGCTACTTGATATGGGGGT
Tet4: TTGGGGTTGGGGTTGGGGTTGGGG
rTet4: UUGGGGUUGGGGUUGGGGUUGGGG
GL: TATGGGGGAGCTGGGGAAGGTGGGATTT
Tet1.5: GGGGTTGGGG
cY: ACCCCCATATCAAGTGACACAGT
cTet4: CCCCCACCCCCACCCCCACCCCCAA

**Mobility shift experiments and competition assays**

Fifty nanograms of 32P-labeled oligo Y was boiled for 3 min in G-DNA formation buffer (50 mM KCl, 10 mM MgCl2, 10 mM Tris-HCl, pH 8.0, 5% glycerol), cooled, and incubated at room temperature for at least 30 min to make G-DNA. Mobility shift experiments were performed as previously described with minor modifications (32). About 2.5 ng of labeled Y(G4) were mixed with desired amount of protein extract or TGPI fractions in binding reaction buffer (10 mM Tris-HCl, pH 7.5, 6% glycerol). One hundred fold (~250 ng) non-specific competitor poly(dI-dC) (Pharmacia) was added to each binding reaction. The total volume of each reaction was 20 µl. After incubation on ice for 20 min, the reaction mixtures were loaded onto a 6% polyacrylamide gel. Electrophoresis was carried out in 0.6 x TBE at room temperature. The gel
was then vacuum-dried, and exposed to X-ray film or a Phosphorimager screen (Molecular Dynamics).

For competition assays, unlabeled competitors were made as follows. Single-stranded DNA (Y) was made by boiling in ddH₂O for 3 min to prevent G-quartet structure formation, and chilling on ice for 30 min. Double-stranded DNAs (Y-cY, Tet4-cTet4) were formed by annealing of single-stranded oligos with their complementary strands at 95°C for 5 min, 65°C for 10 min, and 37°C for 10 min in buffer containing 50 mM NaCl and 10 mM Tris-HCl, pH 7.5. G-DNAs were formed according to the methods described by Sen and Gilbert (34). In mobility shift assays, unlabeled competitors were diluted as indicated and added to the binding buffer before the addition of TGPI and labeled Y(G4) probe. About 50 ng of TGPI and saturating amount (2.5 ng) of labeled probe Y(G4) were used in each binding reaction. Twenty units of RNasin ribonuclease inhibitor (Promega) were included in reactions where RNA oligo (rTet4) was used. Binding reactions and mobility shift experiments were performed as described above.

**SDS-PAGE and silver staining**

SDS-PAGE was performed according to the standard protocol (33). After electrophoresis, protein bands were visualized by silver staining following manufacturer's protocol (Bio-Rad).

**UV cross-linking**

Cross-linking of G-DNA to TGPI was performed *in situ* by exposing a wet 8% mobility shift gel on ice to 254 nm UV light with the gel no more than 5 cm from the UV source. The gel was then exposed to film to reveal the position of the TGPI/G-DNA binding complex. The gel piece containing the complex was excised, denatured in SDS sample buffer for 5 min, and polymerized into the stacking gel of a 10% SDS-polyacrylamide gel. After electrophoresis, the gel was dried and exposed to X-ray film. Molecular weight markers were included to identify the positions of cross-linked bands.

**Tetrahymena cell culture and extract preparation**

*Byzophyllum* cell culture and total protein extract preparation were performed as previously described (32). Briefly, *Tetrahymena thermophila* cells (strain C3V) were grown vegetatively to mid-log phase (2.5 × 10⁵ cells/ml) and harvested. Cells were washed twice with 10 mM Tris-Cl (pH 7.5), resuspended in 5 volumes of TMG buffer (10 mM Tris-Cl (pH 7.5), 1 mM MgCl₂, 10% glycerol and 10 mM β-mercaptoethanol) with protease inhibitors (0.01 mM Leupeptin, 0.01 mM Pepstatin, and 0.1 mM Pefabloc, all from Boehringer-Mannheim), and lysed by addition of one tenth volume of 2% NP-40 (Sigma). The cell lysate was centrifuged at 100,000 × g for 70 min. The supernatant (S100) was aliquoted, immediately frozen in liquid nitrogen, and stored at -70°C.
Purification of TGP1

All purification steps were conducted at 4°C unless otherwise indicated. *Tetrahymena* S100 protein extract (100 ml, -10 mg/ml) was quickly thawed and filtered through a 0.4 μm filter (Costar). The filtrate was then loaded onto a SP-Sepharose column (4.91 cm×14.5 cm, Pharmacia), which had been equilibrated and packed with at least 2 column volumes of chromatography buffer (CB buffer: 10 mM Tris-HCl (pH 7.7), 1 mM EDTA, 0.01% NP-40, 10% glycerol, and 1 mM dithiothreitol (DTT)), at a flow rate of 0.8 ml/min. After the extract had passed through the column, 1.5 column volumes of CB buffer was passed through the column at the same flow rate (0.8 ml/min). The column was then washed with 2 column volumes of CB buffer containing 0.45 M NaCl at a flow rate of 2 ml/min to remove weakly bound proteins. Four column volumes of CB buffer containing 0.8 M NaCl were then applied to elute TGP1 activity. About sixty 3 ml fractions were collected and assayed for TGP1 activity in mobility shift experiments.

Fractions from the SP-Sepharose column containing TGP1 activity were pooled and dialyzed against CB buffer overnight at 4°C. The fractions were then loaded at a flow rate of 0.5 ml/min onto a DE52 column (3.21 cm×4 cm, Whatman), which had been equilibrated with CB buffer. The column was then washed with 1 column volume of CB buffer at a rate of 0.5 ml/min. A linear NaCl gradient (0 to 0.8 M in CB buffer, 2 column volumes) was applied to elute TGP1 activity at a flow rate of 0.4 ml/min. About thirty 2 ml fractions were collected and assayed in mobility shift experiments.

A G-DNA affinity column (~2 ml) was constructed according to a published procedure (20) with minor modifications. Briefly, 200 μg of 3’-biotinylated oligo Y (Midland) were mixed with 1 mg of oligo Y (5 fold) in the presence of 10 mM MgCl₂ and 100 mM KCl. The mixture was boiled for 3 min and then cooled to room temperature for 30 min to allow G-DNA formation. The biotinylated Y(G4) was then mixed with 2 ml of 50% avidin-agarose (Pierce), which had been equilibrated with CB buffer containing 100 mM KCl. After incubation at room temperature for 1 hour, the mixture was poured and packed into a 5 ml disposable column (Bio-Rad). Fractions from the DE52 column containing TGP1 activity were pooled and dialyzed against CB buffer overnight at 4°C. The fractions were then loaded at a flow rate of 0.1 ml/min onto the G-DNA affinity column, which had been equilibrated with CB buffer containing 0.1 M KCl. After the sample passed through, the column was washed with 1 column volume of 0.1 M KCl/CB buffer, and proteins were eluted with a linear KCl gradient (0.1 to 2.0 M) in CB buffer. About thirty 0.2 ml fractions were collected and tested for TGP1 activity in mobility shift experiments.
Protein renaturation

Renaturation experiments were performed using a procedure similar to that described by Hager (35). Briefly, partially purified TGPI (~10 μg) was displayed on a 10% SDS-PAGE, and the gel was then stained with Coomassie-blue. Gel slices containing different protein bands were excised, rinsed with ddH2O, put into different tubes, and crushed into small pieces. One milliliter of elution buffer (50 mM Tris-HCl (pH 7.5), 0.1% SDS, 0.1 mg/ml BSA, 1 mM DTT, 0.2 mM EDTA, and 2.5% glycerol) was then added to each tube. Proteins were eluted from the gel into buffer with agitation for at least 2 hours at room temperature. The mixtures were centrifuged for 2 min at 10,000 rpm. Supernatants were transferred to fresh tubes, mixed with 4 volumes of 100% acetone (-20°C), stored at -20°C for more than 2 hours, and centrifuged for 10 min at 10,000 rpm. Precipitates were rinsed with 0.5 ml of 100% methanol (-20°C) twice, air dried, and dissolved in 5 μl of 6 M guanidine-HCl in dilution buffer (50 mM Tris-HCl (pH 7.6), 20% glycerol, 0.1 mg/ml BSA, 0.15 M NaCl, 1 mM DTT, and 0.1 mM EDTA) for 20 min at room temperature. Solutions were 50-fold diluted with dilution buffer, and proteins were allowed to renature overnight at room temperature. Renatured proteins were tested for TGPI activity in mobility shift experiments.

Kd determination of TGPI/G-DNA binding complex

Affinity purified TGPI fractions were pooled, dialyzed against CB buffer, and concentrated using Centricon concentrators (Amicon). TGPI protein concentrations were estimated by SDS-PAGE and Coomassie blue staining. Concentrated TGPI was then serially-diluted 2-fold and used in mobility shift experiments. Bound and unbound Y(G4) probe was quantitated by a PhosphorImager (Molecular Dynamics). The Kd for TGPI G-DNA binding reaction was estimated from the plot of percentage of bound Y(G4) vs. protein concentration.

Peptide sequencing

Purified TGPI from affinity column was displayed on a 10% SDS-PAGE and electrotransferred to PVDF membrane according to the manufacturer's protocol (Bio-Rad). The 80 kDa band was excised and N-terminal sequencing was attempted (ISU Protein Facility), but failed due to N-terminal blockage, thus internal peptide sequencing was then performed. About 5 μg of TGPI (80 kDa) was eluted from 10% SDS-PAGE gel and acetone precipitated (see the Renaturation section for details). The protein was then subjected to CNBr digestion according to the protocol described by Smith (36). After digestion, peptides were resolved on a 16% SDS-PAGE, and transferred to PVDF membrane. One of the peptides (~35 kDa) yielded a clear peptide sequence: ??GRQSAEG?VG?hYIEV?rFgQYi ('?' indicates that no amino acid had been assigned, while letters in lower case indicate tentative assignments).
Molecular cloning of TGP1 cDNA.

Total RNA was isolated from mid-log phase *Tetrahymena thermophila* cells using TRIzol reagent according to the manufacturer's protocol (Gibco-BRL). First strand cDNA synthesis from total RNA was done by reverse transcription using the M-MLV reverse transcriptase (Gibco-BRL). Based on the internal peptide sequence and with consideration of *Tetrahymena* genetic codon usage (37), two partially degenerate primers were designed: primer 1 [5'-GGTAGA(C/T)A(A/G)(T/A)(C/G)(T/C)GAAGG] corresponding to the peptide sequence GRQSAEG, and primer 2 [5'-AATCTGCA(A/G)ACTTC(A/G)AT(A/G)TA] corresponding to the antisense sequence of peptide YIEV?rF. PCR was done using the first strand cDNAs and primer 1 and primer 2. The PCR product (~55 base pairs) was cloned into TA cloning vector (Invitrogen), and sequenced (ISU DNA Facility).

Both 5' and 3' rapid amplification of cDNA ends (RACE) were performed according to the recommended protocols (Gibco-BRL). For 3' RACE, a TGP1 gene specific primer (primer 3: 5'-GCTGAAGGAAAAGTCGGTGGTCAC) was designed based on the sequence of the 55 bps PCR product. First strand cDNA synthesis was done using the poly(dT)-anchor oligo (Gibco-BRL). Subsequent PCR was performed over the first strand cDNAs using the anchor and primer 3. PCR products were cloned and sequenced. For 5' RACE, the first strand cDNA synthesis was done using a gene specific primer (primer 4: 5'-TTGACCGAATTTCTTAACTTCG), which was based on the DNA sequence obtained in the 3' RACE. After that, the first strand cDNAs were tailed with poly(dC), and purified. PCR was performed over the tailed cDNAs using a poly(dG)-anchor primer and a nested gene specific primer (primer 5: 5'-TAGTGACCACCGACTTTTCC) as primers. PCR products were cloned and sequenced.

Sequence analyses and database searches

GCG (Genetics Computer Group, Inc.) software was used for sequence analyses. BLAST (38) and FASTA (39) database searches were performed at the National Center for Biotechnology Information (NCBI). Secondary structure analysis was done at the European Molecular Biology Laboratory (EMBL).

RESULTS

Identification of TGP1 by UV cross-linking

To identify proteins responsible for TGP1 activity, UV cross-linking experiments were performed. For these experiments, Tet4(G4) probe and a saturating amount of S100 extract were used. The results are shown in Fig. 1. There were no detectable bands in samples which were not exposed to UV-light and which did not contain both the protein extract and probe. In contrast, a band around 95 kDa was observed in UV-exposed samples containing both the protein extract and DNA probe. These results suggested that the 95 kDa band was the cross-linked complex.
Figure 1. UV cross-linking assay to estimate the molecular weight of TGPl. Radiolabeled Tet4(G4) were cross-linked to TGPl in situ. The band containing the cross-linked complex was excised from a mobility shift assay gel and polymerized into a 10% SDS-PAGE gel for size estimation. The -UV/-Pro and +UV/+Pro lanes contained probe alone without and with UV treatment respectively. The +UV/+Pro lanes contained protein/DNA complexes exposed to UV for different amounts of time (in minutes). The main cross-linked complex had an apparent MW of about 95 kDa. A minor complex around 45 kDa was also observed with longer UV exposure times.
between TGPI and DNA probe. Depending on how many Tet4 monomers cross-linked to TGPI, the molecular weight (MW) of the protein can range from 61 kDa to 87 kDa (the Tet4 monomer has a MW about 8.4 kDa). As UV exposure time increased, a minor cross-linked product around 45 kDa was observed. This 45 kDa band could have resulted from a degradation product of the 95 kDa band, or may represent a different protein. Subsequent experiments (affinity purification and renaturation) suggested that this band represents a different G-DNA binding protein.

**Purification of TGPI**

A number of chromatography columns were tested on analytical scales (1-10 ml) to determine their usefulness in purifying TGPI. Three columns (SP-Sepharose, DE52 and G-DNA affinity) were chosen for the purification. The first column used, SP-Sepharose, is a cation-exchange column (Fig. 2A). As shown in the mobility shift assay, TGPI was eluted when the salt concentration reached 0.8 M NaCl. This column resulted in removal of over 90% of non-TGPI proteins (data not shown). The high salt concentration (0.8 M NaCl) needed to elute TGPI activity suggested that TGPI is a basic protein, or may contain basic domains. After the SP-Sepharose column, the TGPI fractions were applied to an anion-exchange (DE52) column (Fig. 2B). TGPI activity was eluted from the column at NaCl concentrations of 0.15 M to 0.5 M.

To further purify TGPI, a G-DNA affinity column was used. After elution from this column, the fractions were subjected to both mobility shift assay and SDS-PAGE silver staining. As shown in Fig. 3A, TGPI activity was eluted between 0.6-1.0 M KCl with a peak at 0.8 M KCl. In the silver staining (Fig. 3B), two major bands (~80 kDa and ~40 kDa) were present in the TGPI peak fraction. However, the 80 kDa band pattern in the silver staining correlated best with TGPI activity shown in the mobility shift assay, suggesting that the 80 kDa band is responsible for TGPI activity. The 40 kDa protein may correlate with the lower shifted bands in the mobility shift assay.

**Renatured ~80 kDa protein contains TGPI activity**

To confirm that the 80 kDa protein was TGPI, renaturation of proteins from SDS-PAGE was performed. The renatured proteins were used in mobility shift experiments to assay for TGPI activity. The results (Fig. 4) showed that the protein renatured from the 80 kDa band had TGPI activity, while no shifted bands were observed for two control proteins (50 kDa and 30 kDa). These results demonstrated that the 80 kDa band contains the TGPI protein. The size of TGPI (80 kDa) is consistent with that predicted by the UV cross-linking experiment. Interestingly, a strong band that shifted to a position lower than the TGPI/G-DNA complex was observed for the renatured 40 kDa protein. To test whether there are any interactions between the 80 kDa and 40 kDa proteins, we mixed both renatured proteins and tested them in the mobility shift assay. No additional shifted bands were observed, suggesting no obvious interactions between these two
Figure 2. SP-Sepharose and DE52 column purification of TGPI. (A) SP-Sepharose column purification. S100 extracts of vegetatively grown *Tetrahymena* mid-log phase cells (2.5 x 10^7/ml) were loaded onto a ~50 ml SP-Sepharose column (Pharmacia). Stepwise NaCl was applied to elute the proteins. Following purification, the fractions were tested for TGPI activity in mobility shift experiments. Most of the TGPI activity eluted from the column when 0.8 M NaCl was applied (fractions 48 to 52). (B) DE52 column purification. The TGPI fractions from SP-Sepharose column were pooled, dialyzed against CB buffer overnight, and loaded onto a ~10 ml DE52 (Whatman) column. Proteins were eluted using a linear salt gradient (0-0.8 M NaCl). TGPI activity was followed by mobility shift assays. Fractions 15 to 20 contained the TGPI activity.
Figure 3. G-DNA affinity column purification of TGPl. The TGPl fractions from DE52 column were pooled, dialyzed, and applied to a 1 ml G-DNA column. A KCl gradient (0-2.0 M) was used to elute proteins from the column. The fractions were tested for TGPl activity by mobility shift assays (A). TGPl activity was eluted at 0.8-1.0 M KCl. The fractions were also tested by SDS-PAGE and silver staining (B) to determine protein components in each fraction. Two major bands (~80 kDa and ~40 kDa) were found in the silver stained gel. The pattern of the 80 kDa band matched that of the TGPl activity.
Figure 4. Renaturation of TGP1 activity from SDS-PAGE gel. Proteins (80 kDa, 50 kDa, 40 kDa, and 30 kDa) from SDS-PAGE gel were allowed to renature according to the procedure described in the methods section. Renatured proteins were tested for TGP1 activity in mobility shift assays (native TGP1 was used in the control lane).
proteins. However, this result does not rule out the possibility of interactions between these two proteins, because the proteins used in the assay were likely to have been only partially renatured.

**TGPl binds specifically to G-DNA**

The DNA binding specificity of TGPl was studied by competition assays using purified TGPl. Oligonucleotides used in the assays are listed in Table 1. The results are shown in Fig. 5. The strongest competitors for TGPl G-DNA binding are GL(G4), Tet(G4) and Y(G4). These three oligos are in the intermolecular parallel G-DNA (G4) form, suggesting that TGPl has high binding affinity for such a structure. Non-G-DNAs, including single-stranded DNA (Y) and double-stranded DNAs (Y-cY, Tet4-cTet4), did not compete for TGPl binding under our assay conditions. Antiparallel G-DNAs (G4' and G'2) showed intermediate competition between those of parallel G-DNAs and non-G-DNAs. G-RNA rTet4(G4) was also tested for competition, and showed much weaker competitive activity than its DNA counterpart Tet4(G4). Hence, our competition results demonstrate that TGPl is a G-quartet structure-specific binding protein that prefers G-DNAs in the parallel form.

**TGPl/G-DNA complex has a Kd of 2.2 \times 10^{-4} \, M**

Purified TGPl was also used to determine the Kd of the TGPl/G-DNA complex. The result of the mobility shift assay using serially-diluted (2-fold) TGPl is shown in Fig. 6A. A supershift band was observed when high concentrations of TGPl were used. With TGPl at low concentration, the supershift band disappeared. This supershift could be a dimer of TGPl binding to a G-DNA substrate. However, because the TGPl fraction used in the assay contains other proteins (mainly the 40 kDa protein), it’s possible that other protein components may be involved in forming such a supershift band. Bands in the mobility shift gel were quantitated and the data were plotted in the graph shown in Fig. 6B. The estimated Kd for TGPl/G-DNA binding is about 2.2 \times 10^{-8} \, M, under our assay conditions (2.5 ng Y(G4) probe, 10 mM Tris-HCl (pH 7.5), 4°C). This Kd value is comparable with those of other G-DNA binding proteins such as G4p1 (17) and G4p2 (18).

**Molecular cloning of TGPl cDNA**

To clone the cDNA coding the TGPl protein, peptide sequencing was performed. No sequence data were obtained from the N-terminal sequencing directly from TGPl, possibly due to N-terminal blockage (data not shown). However, a 35 kDa peptide obtained from CNBr digestion of TGPl yielded an amino acid sequence (??GRQSAEG?VG?hYIEV?rFgQYi).

TGPl cDNA was cloned based on a PCR strategy. Two non-degenerate primers were designed based on the internal peptide sequence of TGPl. Using these two primers, a ~55 base pair fragment was amplified from the first strand cDNAs by reverse transcriptase PCR (RT-PCR).
Figure 5. TGP1 G-DNA binding specificity. In competition assays, 2.5 ng of labeled Y(G4) was used as probe. Appropriate amounts of purified TGP1 were used to keep the Y(G4) probe saturated (excessive probe). Unlabeled competitors were diluted (2, 5, 10, 50 and 100 fold) and used in mobility shift assays, as described in the methods section. The nomenclature of different G-quartet forms was according to Sen and Gilbert (34): G4, G'4 and G'2 denote intermolecular parallel, intramolecular foldback, and antiparallel dimer G-quartets, respectively.
Figure 6. Kd determination of TGPl G-DNA binding. (A) Mobility shift assays were carried out with 2.5 ng Y(G4) probe and 2-fold serial-diluted TGPl (the highest TGPl concentration was about 1 µg per reaction). A supershift band was observed when TGPl of high concentrations was used. (B) Kd estimation. The mobility shift gel (A) was quantitated using a Phosphorimager (Molecular Dynamics). The data were plotted and a Kd of about $2.2 \times 10^{-7}$ M was determined from the graph.
The DNA sequence between the two primers encodes a peptide KVGGH which is consistent with the peptide sequence obtained from direct peptide sequencing. Based on this DNA sequence, the full length sequence of TGP1 cDNA was obtained using 5’ and 3’ RACE.

Translation of the longest open reading frame (ORF) in the TGP1 full length cDNA sequence produces a 725 amino acid long protein (the TGP1 cDNA and protein sequences had been deposited into the GenBank with an accession number AF006380). The sequence of this predicted protein is considered correct and complete for TGP1 for the following reasons: i) this predicted protein sequence contains the peptide sequence obtained from the direct peptide sequencing; ii) the predicted molecular weight for TGP1 from the sequence is 83.2 kDa, which is consistent with the SDS-PAGE result (~80 kDa); iii) the amino acid composition of this predicted TGP1 protein matches the composition data obtained from acid hydrolysis of TGP1 (data not shown).

TGP1 is a novel protein with unusual sequence features

Analyses of the predicted TGP1 protein sequence showed that TGP1 has an unbalanced amino acid composition: it is rich in asparagine (13.9%), lysine (9.4%), glutamine (7.2%), and arginine (7.0%), whereas it has only 1 tryptophan, 2 cystines and 5 methionines. The richness of basic amino acids (lysine, arginine and histidine) results in a predicted pI of 10.58, and likely relates to the DNA binding activity of TGP1. Hydrophobicity plot of TGP1 showed that TGP1 contains two extensively hydrophilic and basic regions. One of the two regions is from residue 476 to residue 603. In this 128 aa region, there are 109 hydrophilic residues, and among them, 40 asparagines (33%, twice the average percentage in the whole TGP1 sequence), 17 glutamines (14%, twice of the average), and more strikingly, 33 basic residues (compared with only 5 acidic residues). Another hydrophilic and basic region is from residue 251 to residue 285. The hydrophilic and basic properties of these regions suggest that they may serve as the DNA binding sites (domains) for TGP1. However, none of the known DNA binding motifs were found in the TGP1 sequence (including the hydrophilic and basic regions).

Comparison of the TGP1 protein sequence with available databases using BLAST and FASTA programs showed that TGP1 does not share significant homology or similarity with any other proteins, including many known G-DNA binding proteins, yet is weakly similar to a number of proteins. An entry with one of the best scores (P=0.02) in the BLAST search is the yeast NUFl protein, which contains coiled-coil structures and is probably a component of the yeast nuclear skeleton (40). The GAP alignment (gap creation penalty: 7; extension penalty: 2) between TGP1 and NUFl over the whole sequence showed a 22% identity. In addition to NUFl, several other proteins with coiled-coil structures, including the tropomyosin α chain of smooth muscle, paramyosin, and myosin regulatory light chain from different species, were returned in the FASTA search. However, a secondary structure prediction program (the PHD program in
A. G4pl x TGPl

46 QLNLVILKNTFTVISLVPSTIDAVVFEVALPLIKDLVASSKIVKSTTTY 95
  .: | | | : | | | | | | : 152 RMDATLALNTAAITQAEIFSNEKSTIQSLESILNSDKYLFSDAVLAFNES
96 RHILFWIDYMQNLLEVSSTKLEINHDLPHEVEKKKAPAGGAADA 145
202 NKLVVRNFRSNTVEANDRQTQLSNNNLRAKEYSDETEKKALTIVLQIH 251
146 AKADEDVSKGKQDGFR 163
  | | | | 252 EKKEKVQGNQNNNFR 268

B. G4p2 x TGPl

226 NNNSRNNEN 234
  | | | 562 NNNSRNNEN 570

Figure 7. Partial sequence alignments between TGPl (GenBank accession No. AF006380) and G4pl (17) (A), and G4p2 (18) (B). The identity (|) and similarities (: and .) between residues are indicated.
EMBL) did not predict similar structures for TGPl. Of note, the BLAST search also returned two yeast G-quartet binding proteins, G4p1 and G4p2. TGPl has a region (152-268 aa) similar (19% identity, 40% similarity) to a region in G4p1 (46-162 aa) (Fig. 7A). TGPl also has several short sequences similar to sequences in G4p2, and one of the short sequences is within the longer extensively hydrophilic region in TGPl (Fig. 7B). However, alignments of TGPl with either G4p1 or G4p2 over the whole sequence resulted in much lower (probably random) similarity.

DISCUSSION

G-DNAs are novel DNA structures whose proposed biological roles include participation in telomere function, recombination, and gene regulation. These putative roles are still unproven and need further exploration. One approach is to identify and study proteins that interact with G-DNA. In this paper, we described the purification, biochemical characterization and molecular cloning of TGPl, the first identified Tetrahymena G-DNA binding protein (32).

This study ultimately identified TGPl as an 83 kDa protein. UV cross-linking experiments suggested that a protein of molecular weight of about 61 to 87 kDa was responsible for the TGPl activity. During a three-column purification process, an ~80 kDa protein was found in the most highly purified fractions. Renaturation experiments demonstrated that the ~80 kDa protein contained TGPl activity. Furthermore, the cloned TGPl cDNA encodes a protein with a predicted molecular weight of 83 kDa. In addition to the 83 kDa TGPl protein, we identified a ~40 kDa protein which copurified with TGPl during the three-column purification. Interestingly, renaturation experiments showed that this 40 kDa protein possessed a G-DNA binding activity distinct from TGPl. Since the N-terminal peptide sequence of the 40 kDa protein was not found in the TGPl sequence (data not shown), the 40 kDa protein is unlikely to be a degradation product of TGPl, but rather represents an additional G-DNA binding protein.

The binding specificity of TGPl was evaluated using the purified protein. The results of these experiments are consistent with a previous study (32). The subtle DNA binding specificity of TGPl distinguishes it from most known G-quartet binding proteins. First, TGPl showed very low affinity for non-G-quartet nucleic acids (i.e., single or double-stranded DNAs), whereas some G-DNA binding proteins are able to bind to these structures. Such proteins include the yeast telomere protein RAP1 which binds to the double-stranded yeast telomeric DNA with much higher affinity (25), and a rat hepatocyte protein qTBP42 which can bind to single-stranded DNAs (41). Second, TGPl has a weaker binding affinity for G-quartet structures formed by RNA, while some G-quartet binding proteins (e.g., a mouse cytoplasmic exoribonuclease mXRN1p (42), two yeast protein G4p1 and G4p2 (17,18)) have high affinity for G-RNA. Third, TGPl has a higher affinity for parallel-stranded G-DNA than for the antiparallel form. In contrast, most known G-DNA binding proteins do not distinguish between these two G-DNA forms. The differences in binding specificity suggest that G-DNA/RNA binding proteins can be
divided into several subfamilies, each with specificity for different G-quartet forms. TGPl could therefore belong to a distinct parallel G-DNA-binding protein subfamily.

Sequence analyses showed that none of the known DNA binding motifs were found in TGPl, including a recently identified Myb-like domain characteristic of some telomere binding proteins (43,44). This result is not surprising in light of the fact that TGPl binds specifically to G-DNA, and has almost no affinity for double or single-stranded DNAs. It is reasonable to predict that novel DNA binding motifs would exist to account for the specific G-DNA binding activity. Consistent with this hypothesis, two extensively hydrophilic and basic regions, which have no similarity to any known DNA/RNA binding motifs, were identified in TGPl. We propose that the hydrophilic and basic regions comprise novel G-DNA binding domains. We further suggest that these novel domains could be shared by other G-DNA binding proteins to some extent. The weak similarities between sequences of G4p2 and the longer extensively hydrophilic region within TGPl lend support to this hypothesis. Since no G-DNA specific domains have been identified in any G-DNA binding proteins thus far, further characterization of these putative G-DNA binding domains might provide valuable information on mechanism(s) by which proteins can specifically recognize the G-DNA structure.

Homology/similarity searches revealed that TGPl is a novel protein with very limited similarity to a number of proteins. The weak similarity between TGPl and NUFl, the yeast putative nuclear skeleton protein, suggests that TGPl may be a component of the nuclear skeleton (matrix) in Tetrahymena nuclei. Nuclear matrix, the insoluble non-chromatin scaffold structure of nucleus, has been thought to be involved in many nuclear events such as chromosome organization, gene localization and gene regulation, through associations with chromosomes (45). The relative abundance of TGPl (about 1.6 x 10^6 molecules/cell, (32)) is consistent with a role for TGPl as a structural protein related to the nuclear matrix. One possibility is that TGPl may bind to telomeres, where G-DNA structures could form (46), linking the telomeres to the nuclear matrix, thus helping to organize the chromosomes. This speculation is supported by data from other species that link telomere proteins to nuclear matrix. For example, the yeast telomere protein RAP1 is associated with the nuclear matrix (47), and interacts with the SIR4 protein (48), which is weakly similar to the human nuclear matrix proteins lamin A and C. In humans, the telomere protein TRF was found to be a component of the nuclear matrix, and the TRF/telomeric DNA complexes are associated with the nuclear matrix network (49).

The proposed function for TGPl in the nuclear matrix is speculative and does not exclude other possible telomere-related functions for TGPl. For example, TGPl may stabilize telomere-telomere interactions that had been observed in many species (50). Such interactions have been suggested to have a role in chromosome separation during mitosis (51,52). A mutation in Tetrahymena telomeric DNA repeat (GGGGTT changed to GGGGTTTT) caused delayed and abnormal separation between the sister chromatids during mitosis (52). The mutant chromatids
became elongated up to twice the normal length when they finally separated, indicating stronger physical association exists between the mutant chromatids, especially at the telomeres (52). However, little is known about how telomeres interact with each other. One of the possible mechanisms is that G-rich strands or 3’ G-overhangs (53–58) of telomeres from different chromosomes can stack together by forming a G-DNA structure (46). If that is the case, protein components such as TGPl could bind to the G-DNA, and stabilize the interaction between sister chromatids. Any alterations in the telomere sequence may change the G-DNA/protein structure, and thus cause problems in chromosome separation.

Preferred TGPl binding to parallel G-DNA suggests that TGPl may also be involved in recombination, since parallel G-DNA structures have long been thought to be involved in recombination processes (7). It was proposed that four G-rich DNA strands from sister chromosomes can bind together to form parallel G-DNA stabilized structures (7). Protein factors that interact with such structures could therefore have roles in recombination. The yeast G-DNA specific nuclease KEM1, which recognizes the parallel G-DNA structure and cuts the single-stranded DNA 5’ to the G-quartet domain in vitro, was proposed to function in such a way in meiotic pairing (20). However, in contrast to KEM1, no nuclease activity was observed for TGPl (data not shown).

All hypotheses regarding the biological functions for TGPl and G-DNA need to be rigorously tested by further studies such as immunolocalization and gene disruption, and these experiments are in progress. These studies will further define the function(s) of TGPl, and should also contribute to the understanding of biological role(s) of G-DNA in general.

ACKNOWLEDGEMENTS

The authors want to thank Scott Schaus for critical reading of the manuscript. This work was funded partly by a research grant from the American Cancer Society to E.H. (NP906). Q.L. was supported in part by fellowships from the MCDB program. This is Journal Paper No. J-17788 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa. Project No. 3064, and supported by Hatch Act and State of Iowa Funds.

REFERENCES

CHAPTER 2: TWO NOVEL TETRAHYMENA G-QUARTET DNA BINDING PROTEINS, TGPl AND TGP3, SHARE REPETITIVE AND HOMOLOGOUS MOTIFS

A paper submitted to Nucleic Acids Research

Quan Lu and Eric Henderson
Department of Zoology and Genetics
Molecular, Cellular and Developmental Biology Program
Iowa State University, Ames, Iowa 50011, USA

ABSTRACT

G-DNA is a four-stranded DNA structure with diverse putative biological roles. We have previously purified and cloned a novel G-DNA binding protein TGPl from the ciliate Tetrahymena thermophila (1). Here we report the molecular cloning of TGP3, an additional G-DNA binding protein from the same organism. Based upon the internal peptide sequence obtained from purified TGP3 protein, we cloned the complete cDNA and genomic DNA of the TGP3 gene. The TGP3 cDNA codes for a 365 amino acid long protein. Homology search of the deduced TGP3 protein sequence revealed that TGP3 is homologous to TGPl with 34% identity and 44% similarity. Sequence analyses showed that both proteins are very basic with similar predicted pIs (~10.5), and contain an extremely hydrophilic region. Intriguingly, the sequences flanking the hydrophilic regions in both proteins share significant homology with each other. Thus, TGPl and TGP3 have a similar sequence pattern containing two repetitive regions flanking an extensively hydrophilic region. We suggest that this unusual sequence pattern may constitute a novel G-DNA-specific binding domain.

INTRODUCTION

G-DNA, also known as G4, G-quartet or G-tetraplex DNA, is a family of four-stranded DNA structures characterized by a novel motif called the G-quartet (2,3). In a G-quartet, four guanine molecules from each DNA strand lie in a plane with each guanine forming G:G hydrogen bonds with its two neighbors. Although the existence of G-DNA in vivo remains undetermined, many guanine-rich (G-rich) sequences of biological relevance can readily form G-DNAs in vitro under physiological conditions. These sequences include most telomeric DNA (4-6), immunoglobulin switch region (7), microsatellite sequences (8), and a few gene promoters (9-11). The high stability of G-DNA has evoked the suggestion that this degree of stability may preclude the dynamic properties required for biological activity. However, recent identification of

1 To whom correspondence should be addressed
several proteins with G-DNA unwinding activities (12-14) argues in favor of a functional role for G-DNA in vivo.

Moreover, inductive evidence is accumulating which implies that G-DNA plays roles in many diverse biological processes. First, G-DNA has implied function(s) at telomeres, the terminal structure of eukaryotic chromosomes (15). Two telomere-binding proteins, the yeast RAP1 protein (16) and Oxytricha TEBP beta subunit (17), were found to facilitate the formation of G-DNA in vitro. G-DNA formation at telomeres may regulate the activity of telomerase, the enzyme responsible for the regulation of telomere length, as suggested by a study showing that telomerase was inhibited by G-DNA primers (18). Second, G-DNA may have roles in recombination. The yeast KEM1 protein, which is essential for meiotic recombination, was found to be a G-DNA specific nuclease (19,20). LR1, a B-cell specific protein binding to the immunoglobulin switch region (which is guanine-rich), was found to have G-DNA binding activity (21). Third, G-DNA may play a role in gene regulation. This is suggested by studies identifying G-DNA binding activities in proteins known to be involved in the control of gene expression, including eukaryotic topoisomerase II (22) and transcription factor MyoD (23). Finally, G-DNA may play other roles in vivo, as novel G-DNA binding proteins, including a hepatocyte chromatin protein QUAD (24), two yeast proteins G4p1 and G4p2 (25,26), and Tetrahymena TGP1 and TGP2 proteins (1,27,29), have been identified.

The ciliate Tetrahymena thermophila is a good source of G-DNA, since its macronuclear genome is fragmented and thus contains a large proportion of telomere DNA which is guanine-rich T2G4 repeats (28). Three G-DNA binding activities (TGP1, TGP2, and TGP3) had been identified from Tetrahymena total protein extract (1,27,29). While TGP2 was found to contain dihydrolipoamide dehydrogenase activity (29), TGP1 was shown to be a novel protein with no significant homologs in available databases (1). During TGP1 purification, an additional G-DNA binding activity was identified which appeared to copurify with TGP1. A 40 kDa protein was found to be responsible for this activity. This protein was referred to as TGP3 (Tetrahymena G-DNA binding protein 3). In this paper we report the molecular cloning of TGP3. Sequence analyses showed that TGP3 is homologous to TGP1, and both proteins share an intriguing sequence pattern which may constitute a novel putative G-DNA binding domain.

MATERIALS AND METHODS

Protein purification

TGP3 was purified using essentially the same protocol previously described for TGP1 purification (1). Briefly, S100 total protein extracts were made from Tetrahymena thermophila (strain C3V) cells grown to mid-log phase. Three columns were used for TGP3 purification, including an SP-Sepharose column, a DE52 column and a G-DNA affinity column. After each column, the fractions were tested for TGP1 G-DNA binding activity by mobility shift assays.
Peptide sequencing

Partially purified TGP3 (~5 μg) was resolved by 10% SDS-PAGE, and transferred to PVDF membrane. The protein band was excised and used for N-terminal peptide sequencing (ISU Protein Facility). For internal peptide sequencing, ~5 μg of TGP3 protein transferred to PVDF membrane were digested with CNBr according to Smith (30). Digested peptides were separated by 16% SDS-PAGE, and transferred to PVDF membrane. Each peptide band was excised and subjected to N-terminal sequencing.

Tetrahymena DNA and RNA preparation

Total genomic DNA was prepared using a standard protocol (31). RNA was extracted using TriZol reagent according to the manufacturer’s recommended protocol (Gibco-BRL). *Tetrahymena thermophila* (strain CU428) cells grown vegetatively to mid log-phase were used for both RNA and DNA extraction.

cDNA and genomic DNA cloning

Rapid amplification of cDNA ends (RACE) was accomplished with kits from Gibco-BRL. For 3' RACE, first strand cDNA synthesis was performed using a poly(dT)-anchor oligo (Gibco-BRL). Subsequent PCR was performed with the first strand cDNAs using the anchor oligo and a partially degenerate oligonucleotide TGP3U [5'-TGCAGAAC(T/C)AACAA(T/C)TA(T/C)AGAA AA] corresponding to the peptide cRTNNYRK. The predominant amplified product was cloned and sequenced. For 5' RACE, two gene specific primers 3R1 (5'-TGTTAGTGTTGTTGTTGTT GC), and 3R2 (5'-AGCTTAGTGGAAATCTCTTAGGC) were synthesized based on sequences obtained by 3' RACE. Primer 3R2 was used in the first round PCR, whereas 3R1 was used as a nested primer in the second round of PCR. The predominant PCR product was cloned and sequenced (ISU DNA Facility).

TGP3 genomic DNA was amplified using two gene specific primers termed 3U (5'-TAACAACTAAGTCTCTCCTC) and 3R (5'-ATTCACTCATTGCTTAGTGGC). A 1.6 kb PCR product was purified and sequenced.

Sequence analyses

The GCG (Genetics Computer Group, Inc.) software package was used for sequence analyses. Blast searches were performed over the World Wide Web (WWW) at the National Center of Biotechnology Information (NCBI). The Boxshade program was used to process multiple sequence alignments.
RESULTS

Identification and purification of TGP3

During TGP1 purification, a 40 kDa protein was found to copurify with TGP1 (1). Renaturation experiments demonstrated that the 40 kDa protein accounts for an additional G-DNA binding activity which was referred to as TGP3. Because of its copurification with TGP1, TGP3 was purified using essentially the same purification procedure as that for TGP1. The purification included three chromatographic steps: SP-Sepharose, DE52, and G-DNA affinity chromatography. Fractions from the G-DNA column contained two predominant proteins (80 kDa TGP1 and 40 kDa TGP3) as revealed by silver-stained SDS-PAGE (1).

Molecular cloning of TGP3 cDNA and genomic DNA

To clone the 40 kDa TGP3 protein, direct peptide sequencing was performed. Initially, direct N-terminal peptide sequencing was performed with purified TGP3. However, no clear sequence data were obtained, probably due to N-terminal blockage of the purified TGP3 protein. Internal peptide sequencing was thus attempted. TGP3 was digested with CNBr resulting into three major peptides (data not shown). Each of the peptides was then excised and subjected to N-terminal sequencing. One of the peptides yielded a clear sequence (RTNNYRKNQNQQRKNN). Based on this peptide sequence and with consideration of the *Tetrahymena* genetic codon usage (32), a long partially degenerate primer, TGP3U, was designed. This primer was used directly for 3' RACE (the other primer for the amplification was the anchor oligonucleotide provided in the RACE kit). A short PCR product around 50 base pairs (bps) was amplified and sequenced. Its deduced amino acid sequence contains part of the peptide sequence obtained directly from internal peptide sequencing, indicating that the PCR product resulted from specific amplification of putative TGP3 cDNA. Using the sequence information generated from the 3'-RACE, 5'-RACE was subsequently performed. The full length TGP3 cDNA was thus obtained. The deduced amino acid sequence (Fig. 1) of the cDNA contains the peptide sequence obtained from direct peptide sequencing, and the predicted molecular weight of the deduced protein is about 40.2 kDa, consistent with the protein size observed in SDS-PAGE (1). Taken together, these data suggest that the cloned TGP3 cDNA sequence is likely to be complete and correct. In addition to cloning of the cDNA, TGP3 genomic DNA was also cloned based on a PCR strategy. Two primers (each from the 5' or 3' end of the cDNA sequence) were used in PCR amplification of *Tetrahymena* genomic DNA. A predominant 1.5 kb band was amplified and sequenced. The genomic DNA contains 5 exons and 4 introns. The sequences of the exons match exactly the TGP3 cDNA sequence, confirming the authenticity of the obtained TGP3 cDNA sequence. All the introns contain consensus eukaryotic splicing sites and are AT-rich. The intron lengths range from 42 bps to 187 bps.
MAEQVTNNQVSPQGEQKPKVQKRPRTVQLESKDPVKIADLQPRNNNSNFV
GKVIEVQVLEKGNNKQGNPRKFLKGLIGDDTWVVRFDLAVKNDVFKVDD
VVSFDKAMNKNKGDGHYIEVKRFQGKEYITNGSIAAVKTDNNISTKIIPP
LPEGEKKQKIIKSKNNKQNNNGKVNNNNNDNDNTRKKNNRTNY
RKONNNQORNNNNNSNNNNTNNQSQNYKQKQKPRFHQALLTPFKNKL
PGQNGQTGKVFQDQSYQKTKNKREATFFKGKVADDTANINFDFMMKEK
TISEGDMVIFTPNNKVSGETGHYIVGVGKNYSVLFHNEKQININEQENKS
SIEYAISSSNNATKQ

Figure 1. Deduced TGP3 protein sequence. The peptide sequence obtained from internal peptide sequencing is underlined. The putative nuclear localization sequences (NLS) are dot-underlined. The sequence of TGP3 cDNA has been deposited into the GenBank (Accession number: AF136448).
TGP3 is homologous to TGP1

Sequence analyses showed that TGP3 has several interesting sequence features. First, TGP3 is rich in basic amino acid residues and has a predicted pI of 10.67. Second, TGP3 contains a very hydrophilic region (residues 154 to 240) which is also rich in asparagines and basic residues. This hydrophilic region was predicted to be capable of forming a coiled-coil structure. Third, TGP3 contains two putative nuclear localization signals (NLS). Notably, all three sequence features are also shared by TGPl (1). For example, TGPl has a pI about 10.5, has a hydrophobicity plot strikingly similar to that of TGP3 (Fig. 2), and contains putative NLS sequences. However, even though TGPl and TGP3 share these sequence features, they do not contain any known DNA/RNA binding motifs including the telobox motif found in many telomeric DNA binding proteins (33).

Sequence comparison revealed that TGPl and TGP3 are homologous to each other but share no significant similarities with any other proteins in available databases, suggesting that they are indeed novel proteins. Sequence alignment between TGPl and TGP3 showed that TGP3 is homologous to the 3' end of TGPl but lacks about 250 amino acids of the 5' end of TGPl (Fig. 3A). The sequence identity between the two proteins over the alignment regions is 34% and the level of similarity is 44%. The alignment also creates a long artificial gap region in the TGP3 sequence. The hydrophilic regions immediately follow this gap. The alignment between these regions is of very low complexity (most identical residues are asparagines), and, therefore, probably only reflects the hydrophilic and asparagine-rich nature of the sequences. In contrast, the alignment between regions flanking the gap and the hydrophilic regions is of high complexity, and the identical residues are often in clusters. For the purpose of clarity in the following discussion, we designate the sequence upstream of the hydrophilic region (and the gap) as region A, the downstream sequence as region B.

Repetitive and homologous motifs in TGP sequences

When we used region A of TGP3 sequence in a BLAST search, we found that it is homologous not only to region A of TGPl (which is expected from the sequence alignment shown in Fig. 3A) but also to region B of TGPl. Using region B in a BLAST search, we found a similar situation. This cross-homology between A and B regions suggests that the four regions (two regions from each protein) are homologous to each other. Subsequent multiple sequence alignments revealed that similarities exist among the four regions (Fig. 3B). There are 16 identical amino acid residues among all the four regions, and 22 additional sites identical among at least 3 out of the four regions. Even though these identical sites constitute only about 20% of the whole sequence in the alignment, most of these sites tend to be in clusters suggesting that the similarities are functionally significant. In summary, the sequence alignments revealed that TGPl and TGP3 share a similar sequence pattern containing two novel repetitive, homologous motifs.
Figure 2. Hydrophobicity plots of TGP1 and TGP3 share strikingly similar profiles. Both proteins contain a very hydrophilic region. The hydrophilic regions contain large number of asparagine residues and are also rich in basic amino acid residues.
Figure 3. TGP1 and TGP3 are homologous. (A) Sequence alignment (GAP creation penalty: 10, extension penalty: 2) between TGP3 and TGP1. (B) Sequence alignment among four domains (two from TGP1, 2 from TGP3) flanking the hydrophilic regions. Identities and similarities are indicated by differently-colored shadings. The program Boxshade was used for the alignment processing.
flanking an extensively hydrophilic and basic region. No proteins with a significantly similar pattern were found in BLAST homology searches suggesting that this pattern may represent a novel protein domain or motif, possibly involved in G-DNA binding.

DISCUSSION

G-quartets are novel nucleic acid structures for which definitive evidence of their existence and function in vivo remain elusive and need further exploration. We have previously identified and cloned a novel Tetrahymena G-quartet DNA binding protein TGPl. In this paper we reported the molecular cloning of an additional Tetrahymena G-quartet DNA binding protein TGP3 which copurifies with TGPl during a three-column purification process. Intriguingly, TGP3 and TGPl share significant homology and a novel sequence pattern. To our knowledge, this represents the first example of homology among all known G-quartet binding proteins.

The homology between TGPl and TGP3 implies that TGP homologues may exist in other organisms. However, we found no significant homologs of TGPs in available protein databases, nor did we find proteins with similar sequence motifs shared by TGPl and TGP3. It is somewhat disappointing not to find TGP homologs in yeast and C. elegans, the genomes of which have been completely sequenced. Nevertheless, many ciliate proteins have no known yeast homolog, including the Oxytricha telomere-end binding protein α and β subunits (34), and Tetrahymena telomerase-associated proteins p80 (whose human homolog has been identified (35)) and p95 (36). It is possible that proteins with structure similar to that of TGPs may exist in other organisms. That is the case for the Oxytricha telomere-end binding protein which has no apparent sequence homologs but contains structural motifs called oligonucleotide/oligosaccharide folds shared by many other proteins (37). It may be more productive to initially search for TGP homologs from other ciliates such as Oxytricha and Euplotes. Such homolog cloning work is in progress. Identification of TGP homologs from these organisms will help us define the putative G-DNA binding domain and search for homologs in higher organisms including humans.

The structural basis of G-DNA/protein interactions is not known. Since G-quartet structures are distinctly different from double or single-stranded DNA, it is likely that proteins which specifically bind to G-quartets would have distinct structural motifs. In this sense, it is not surprising to find that TGPl and TGP3 contain no known DNA/RNA binding motifs but share an interesting amino acid sequence pattern. This sequence pattern is composed of two repetitive motifs flanking an extensively hydrophilic and asparagine-rich region. We speculate that this sequence pattern may constitute a novel G-quartet specific binding domain or motif. We further propose a model in which the two flanking repetitive regions may specifically recognize and bind to the G-quartet DNA, whereas the hydrophilic region which has much lower sequence complexity increases the binding strength by electrostatic interaction with the DNA substrate. However, it is currently not clear whether the whole sequence pattern is required for specific G-
Figure 4. A model for TGP/G-DNA binding. The two homologous motifs of the TGP proteins specifically recognize and bind to G-DNA (probably at the major grooves). The hydrophilic and basic region between the homologous motifs strengthens the protein/G-DNA binding through electrostatic interaction with the G-DNA.
DNA binding. It is possible that one of the repetitive motifs alone may be enough for the binding. Detailed domain analysis by serial deletion should identify a minimal domain accounting for the G-DNA binding. Subsequent X-ray crystallography or other high resolution studies on the minimal domain/G-DNA binding complex will provide valuable insights about the structural basis of G-DNA/protein binding.

The biological functions of TGPl and TGP3 remain to be determined. Preliminary gene disruption experiments have suggested that TGPl and TGP3 may play a role in micronuclear division (Q. Lu and E. Henderson, unpublished data). One possibility is that the proteins function in micronuclear division through binding to G-quartet DNA formed between telomeres of sister chromatids. The homology between TGP3 and TGPl suggests that the proteins may have similar functions in vivo, and furthermore there could be certain level of functional redundancy between these two proteins. Indeed, we found that TGP3 activity increases significantly in TGPl knockout cells, suggesting that TGP3 may compensate for the loss of TGPl by increasing its expression. An ongoing double gene disruption experiment and other functional analyses may yield significant information about the biological roles of the proteins and the G-DNA structure in general.

ACKNOWLEDGMENTS

The authors want to thank Dr. Sang-Gu Kang and Scott Schaus for insightful discussion on the manuscript. This is Journal Paper No. J- of the Iowa Agriculture and Home Economics Experiment Station, Ames, IA. Project No. 3064, and supported by the Hatch Act and State of Iowa Funds.

REFERENCES
CHAPTER 3: FUNCTIONAL ANALYSES OF TWO NOVEL TETRAHYMENA G-DNA BINDING PROTEINS: A ROLE IN MICRONUCLEAR DIVISION

A paper to be submitted to Molecular and Cellular Biology

Quan Lu and Eric Henderson
Department of Zoology and Genetics
Molecular, Cellular and Developmental Biology Program
Iowa State University, Ames, Iowa 50011

ABSTRACT

G-DNA is a family of four-stranded DNA structures whose *in vivo* roles remain to be determined. We have previously purified and cloned two G-DNA specific binding proteins (TGP1 and TGP3) from the ciliate *Tetrahymena thermophila* (23, 24). TGP1 and TGP3 are novel proteins but share significant homology with each other. In this paper we report the functional analyses of the two proteins. We have generated *Tetrahymena* macronuclear gene knockout (KO) strains (TGP1KO and TGP3KO) separately for each of the two genes. Southern analysis showed that the macronuclear copies of each gene were completely disrupted. Mobility shift assay showed that the corresponding G-DNA binding activity for each of the proteins was abolished in the KO strains, confirming the complete disruption of the genes in the macronucleus. Interestingly, mobility shift assay also revealed an increased TGP3 activity in TGPIKO cells, implying that these two proteins may share some aspects of biological functions. Growth analysis showed that both KO cells grow at near wild-type rate indicating that neither of the genes is essential for cell growth. Nevertheless, nuclear staining analysis revealed that both KO cells have an increased occurrence (more than two fold) of extra micronuclei, suggesting a faulty control of the micronuclear division in the KO cells. In addition to the gene knockout experiments, we also did nuclear fractionation experiment, demonstrating that both TGP1 and TGP3 localize mainly to the nuclei. Based on these preliminary data, we propose a model in which TGP proteins coordinate to function in micronuclear division through binding to the G-DNA structure formed between telomeres of two sister chromatids.

INTRODUCTION

G-quartets are novel nucleic acids motifs characterized by guanine-guanine hydrogen bonding (10, 13, 28). In a G-quartet, four-guanine molecules lie in a plane with each guanine forming guanine-guanine hydrogen bonds with its two neighbors. Under appropriate salt
conditions, several continuous G-quartets can stack upon each other to form a four-stranded DNA structure G-DNA (also known as G4, G-tetraplex or G-quadruplex DNA).

G-DNA has been proposed to play roles in many biological processes. One of these is at the telomeres, the nucleoprotein complex at the end of eukaryotic chromosomes (34). In most organisms, telomeric DNA is composed of simple tandem repeats. These short repeats are usually G/C rich (e.g., TTGGGG/AACCCC in *Tetrahymena thermophila*). The 5' to 3' strand of telomeres (toward the end of chromosome) is guanine-rich and extends beyond the double-helix part to form a 3' overhang. Numerous studies have demonstrated that these guanine-rich telomeric repeats can adopt G-DNA structures in vitro under near physiological conditions (13, 32), suggesting that G-DNA may exist at telomeres. Consistent with this, two telomere-binding proteins, the yeast RAP1 protein (9) and the *Oxytricha* telomere-end binding protein β subunit (6), were found to be able to facilitate the formation of G-DNA in vitro. The potential G-DNA formation at telomeres may play a role in telomere length regulation (33) and telomere-telomere association (11).

Another role that G-DNA may play is in recombination. Many recombination hot spots such as the immunoglobulin switch region (29) are guanine-rich. In vitro study showed that these sequences can form G-DNA, promoting the speculation that G-DNA can bring together four homologous chromatids during meiotic recombination (27). Supporting this, LR1, a B-cell specific protein binding to the immunoglobulin switch region, was found to have G-DNA binding activity (4). During recombination, strand exchange must occur, and this exchange process requires the participation of nuclease activity. Two proteins known to be involved in recombination, the yeast KEM1 protein (21, 22) and the eukaryotic topoisomerase II (3), were found to contain G-DNA nuclease activity cutting the single-stranded region adjacent to the G-DNA domain.

G-DNA may also play roles in other biological processes. Guanine-rich promoters which can form G-DNA in vitro have been found in several genes (5, 17, 20), and a transcription factor MyoD was found to contain G-DNA binding activity (30), suggesting that G-DNA may exist at the promoters and play a role in gene regulation. G-DNA may as well play some novel biological roles since novel G-DNA binding proteins, including a hepatocyte chromatin protein QUAD (31), two yeast proteins G4p1 and G4p2 (7, 8), and *Tetrahymena* TGP1 and TGP3 protein (23, 24), have been identified.

Putative biological roles of G-DNA are diverse but remain to be definitively determined. To further explore the biological relevance of G-DNA, we studied two G-DNA binding proteins in the ciliate *Tetrahymena thermophila*, an organism rich of telomeres (T2G4 repeats). We have identified, purified and cloned two G-DNA binding proteins from this organism. These proteins, namely TGP1 and TGP3, are novel proteins with no significant homologues in available databases but share significantly homologous motifs with each other (23, 24). In this paper we investigate
the biological functions of these two proteins by disrupting the genes in the somatic macronucleus. Each of the TGP genes has been completely disrupted as was demonstrated by southern blot analysis and mobility shift assay. Both gene knockout strains (TGP1KO and TGP3KO) grow at near normal rate suggesting that neither of the genes is essential for cell growth and survival. However, nuclear staining analysis revealed that, in both KO cells the percentage of cells containing multiple (>2) micronuclei is much higher than that in the wild-type cells, suggesting a faulty control of the micronuclear division in the KO cells. In addition to the gene knockout analysis, we also determined by nuclear fractionation that both TGP1 and TGP3 localize mainly in the nuclei. To account for these data, we propose a model to explain how these G-DNA binding proteins may function in micronuclear division.

MATERIALS AND METHODS

_Tetrahymena_ cell culture

Cells were cultured in 2% PPYS media at 30°C with vigorous shaking (125 rpm). Usually cells were not allowed to exceed concentration of 5 x 10⁵ cells/ml. Cells were transferred to fresh media at an 1:1000 dilution. Stock cultures were maintained in either 1% PPYS for 1 month or soybean media (1 soybean autoclaved in 10 ml ddH₂O) for 6 months.

Construction of gene knockout (KO) vectors

p4T2-l containing the neo (neomycin coding gene) cassette was obtained from Dr. M. Gorovsky at the University of Rochester. pGEXT-TGP1 and pGEXT-TGP3 are pGEXT-easy vectors (Promega) containing genomic DNA of TGP1 and TGP3 genes respectively. To construct gene knockout (KO) vectors, pGEXT vectors were amplified by inside-out polymerase chain reaction (PCR). The primers used are: for TGP1, 1KOU (CCGCTCGAGGCTAAGGTAGCTGTCATTC) and 1KOR (CGGGATCCCACTGCTGATCCAAGCTG); and for TGP3, 3K0U (AACTCGAGATAATTCCTCCTCTTCCTG) and 3K0R (TTCCCCGGTTATCTGTTTTAACAGCGGC). The PCR products and the plasmid p4T2-1 were digested by XhoI and BamHI (or Xmal for TGP3) and ligated to create KO vectors as shown in Figure 1. For use in _Tetrahymena_ transformation, the KO vectors were linearized by ApaI digestion, phenol/CIA extracted twice, ethanol precipitated, air dried and dissolved into ddH₂O at about 2 μg/μl. Four μg of DNA was used for each transformation.

Macronuclear (Somatic) DNA transformation

Gold particle-mediated biolistic gun protocol was used for transformation. This protocol was provided by Dr. M. Gorovsky at the University of Rochester. Linearized KO vector DNA was coated to gold particles as follows. Briefly, 5 μl of DNA (2 μg/μl) was mixed with 25 μl of 1.0 μm gold particles (Biorad, pretreated and stored at -20°C in 50% glycerol at a concentration of 60
mg/ml). Ten μl of 0.1 M Spermidine (Sigma) and 25 μl of 2.5 M CaCl₂ were added to the mixture and briefly vortexed. The mixture was then vortexed at setting 4 at 4°C for 10 min. After that, the gold particles were washed with 70% ethanol and 100% ethanol each for once, and resuspended in 20 μl of 100% ethanol. Ten μl of the resuspended DNA-coated gold particles was loaded onto a flying disk (Biorad) and air dried. *Tetrahymena* cells (strain CU428) were grown to mid-log phase and starved in 10 mM Tris-HCl (7.5) at 30°C with shaking for about 15 hr. Starved cells (1 × 10⁷ cells) were washed, resuspended in 1 ml of 10 mM HEPES buffer (7.5), and spread onto a moist filter paper (Whatman) and bombarded with DNA-coated gold particles at 900 psi using a biolistic PDS-1000/He particle delivery system (Biorad). After shooting, cells were immediately resuspended into 50 ml 2% PPYS media, cultured at 30°C for 4 hr to allow cell recovery, and plated into 96-well microtiter plates at 150 μl/well. Paromomycin (Sigma) was immediately added into the wells at a final concentration of 120 μg/ml. After incubation in a moist chamber for at least 3 days, wells with actively growing cells were counted and replicated to plates with fresh 2% PPYS media containing paramomycin (200 μg/ml).

**Genomic DNA extraction**

*Tetrahymena* genomic DNA was extracted according a standard protocol (1). Essentially, about 1 × 10⁷ mid-log phase cells were pelleted, washed with 10 mM Tris-HCl (7.5), and resuspended in 100 μl warm (37°C) NDS solution (2% SDS, 0.5M EDTA, 10 mM Tris-HCl, pH 9.5) and 200 μl pronase (10 mg/ml), incubated at 55°C overnight (usually 14 hr). The samples were then mixed with 300 μl ddH₂O, extracted twice with phenol / CIA (1:1), precipitated with cold 100% ethanol. The pellets were rinsed with 70% ethanol at least twice, air-dried and resuspended into 100 μl ddH₂O. The samples were further treated with RNase (final concentration of 10 μg/ml, 37°C, 4 hr), extracted twice with phenol/CIA and ethanol precipitated. DNA concentrations were determined using a Beckman DNA photospectrometer.

**Southern blot analysis**

About 15 μg of genomic DNA were digested with appropriate enzyme(s) at 37°C overnight, purified by phenol/CIA extraction and ethanol precipitation, and separated onto 0.8 or 1.5% agarose gel. The gels were depurinated, denatured and blotted onto MagnaGraph nylon membrane. Membrane blots were probed with TGP1 or TGP3 gene fragment (³²P-labeled using a random-primer labeling kit from Promega). In the case of telomere length assay, the blots were probed with pTrel plasmid DNA which contains cloned *Tetrahymena* telomere sequences. Hybridization was done according to a standard protocol (2). The blots were exposed to X-ray films or to Phosphoimager screen (Molecular Dynamics).
Growth Analysis

Tetrahymena cells were grown to mid-log phase ($2 \times 10^5$ cells/ml) at 30°C with shaking. The mid-log phase grown cells were used to inoculate 50 ml 2% PPYS media at starting cell density of 200 cells/ml. Cells were cultured at 30°C with constant shaking (125 rpm). At different time points during the culturing, 100 µl of cell culture was taken out to count the cell number using a Coulter counter (Coulter Electronics). The cell numbers at different time points were then plotted into a growth curve. Cell doubling time was estimated by curve fitting (Microsoft Excel). Usually cell culturing and counting were repeated at least twice to minimize sample variations.

Mobility shift assay

Mobility shift assays were performed as previously described (24). Fifty nanograms of $^{32}$P-labeled oligo Y was boiled for 3 min in G-DNA formation buffer (50 mM KCl, 10 mM MgCl$_2$, 10 mM Tris-HCl, pH 8.0, 5% glycerol), cooled, and incubated at room temperature for at least 30 min to make G-DNA. About 2.5 ng of labeled Y(G4) were mixed with desired amount of protein extracts in binding reaction buffer (10 mM Tris-HCl, pH 7.5, 6% glycerol). One hundred fold (~250 ng) non-specific competitor poly(dl-dC) (Pharmacia) was added to each binding reaction. The total volume of each reaction was 20 µl. After incubation on ice for 20 min, the reaction mixtures were loaded onto a 6% polyacrylamide gel. Electrophoresis was carried out in 0.6×TBE at room temperature. The gel was then vacuum-dried, and exposed to X-ray film or a Phosphorimager screen (Molecular Dynamics).

Hoechst nuclear staining

Mid-log phase grown cells were washed and resuspended in 10 mM Tris-HCl (7.5) at a concentration of $10^5$ cells/ml. Cell were then fixed with the addition of one hundredth volume of 37% formaldehyde at room temperature for 1 min, and stained with 10% of Hoechst DNA dye (Sigma) solution (10 mg/ml) for 10 min. Nuclei were visualized using an Olympus BH-2 fluorescence microscope with 40x lens. The images were captured by a CCD camera, and processed in the program NIH Image.

Nuclear fractionation

Tetrahymena nuclear fractionation was performed according to a standard protocol described by Higashinakagawa (14). Briefly, mid-log phase cells were harvested and resuspended in 10 volumes of ice-cold 0.25 M sucrose and 10 mM MgCl$_2$. NP40 (Sigma) was then added to the cells at a final concentration of 0.2%. The mixture was vigorously pipetted (usually 50 times) on ice until 80-90% of cells were lysed (check under microscopy). Solid sucrose was added to the lysed cell solution to a final concentration of 2.1 M. The mixture was then stirred vigorously (at
maximum setting) at 4°C for 40 min, centrifuged at 50,000 rpm (Ti70.1 rotor, Beckman) for 2 hr. The supernatant was taken out and dialyzed against the Tris buffer (10 mM Tris-HCl (7.5), 1 mM MgCl₂ and 5% glycerol with protease inhibitors added) at 4°C overnight, and stored at -80°C as the cytoplasmic fraction. The pellet containing mostly nuclei (check by methyl green staining) was rinsed with the Tris buffer twice, resuspended in the buffer, and lysed by the addition of 1/10 volume of 2% NP40 and maximum stirring at 4°C for 30 min. This lysed mixture was then centrifuged at 50,000 rpm for 1 hr. The supernatant was stored at -80°C as the nuclear fraction.

RESULTS

Macronuclear transformation of knockout (KO) vectors

To knock out the TGP genes, gene knockout vectors (pTGPlKO and pTGP3KO for TGP1 and TGP3 respectively) were constructed (Fig. 1). Both vectors use a neo (neomycin) expression cassette which contains the promoter of *Tetrahymena* histone H4 gene, a neo gene from Tn5, and a terminator from *Tetrahymena* β-tubulin gene. The neo cassette confers paromomycin resistance when expressed in *Tetrahymena* cells (16). In both TGPlKO and TGP3KO vectors, the neo cassette was inserted into the exons (to ensure that the neo cassette would not be excised during transcription of TGP genes) near the middle of the TGP genes. KO vectors were transformed into starved *Tetrahymena* CU428 cells using a gold particle mediated transformation protocol. Cells were selected for paromomycin resistance. Several dozens of transformants were obtained for both TGPl and TGP3 disruption vectors. Two transformants from each group (1KO, 1KO; and 3KO, 3KO) were randomly selected for further analyses.

Complete macronuclear gene disruption in TGPlKO and TGP3KO cells

*Tetrahymena* macronucleus is not a diploid nucleus (the micronucleus is) but rather contains an average of ~45 copies of each gene (26). In a typical gene transformation, initially gene disruption occurs only in very few (most likely one) copies of the gene through homologous recombination. However, the macronucleus divides amitotically (distribute gene copies randomly to daughter cells) resulting in a phenomenon known as the phenotypic assortment (26). Under selection pressure (such as drug selection), the initial one disrupted gene copy tends to be assorted to homogeneity. If the gene is not essential for survival, a complete gene disruption will be achieved. On the other hand, an essential gene can only be partially assorted.

The selected KO transformants were cultured in 500 µg/ml paromomycin media for at least 200 generations to maximize phenotypic assortment (usually 50 generations is needed for maximum gene assortment). Genomic DNA was then isolated from the wild-type and the KO cells and used in southern blot analysis to determine the extent of gene disruption. The southern results are shown in Figure 2. In the analysis of TGPl knock out (Fig. 2A), a TGPl-gene specific
Figure 1. Schematic diagram of knockout vector construction and transformation.
Figure 2. TGPI (or TGP3) gene in the macronucleus was completely disrupted. (A, C) somatic genomic organization of TGPI (or TGP3) gene. (B, D) Southern blot analysis of TGPI (TGP3) knockouts. wt, wild-type Tetrahymena genomic DNA; 1KO and 1KOB, two of the TGPI knockout stains; 3KO and 3KB, two of the TGPI knockout stains. The TGPKO cells were grown in paromomycin-containing media for over 300 generations Genomic DNA was extracted from the cells, digested with restriction enzymes and hybridized with TGP gene-specific probes.
probe was used to hybridize EcoRI or EcoRV digested genomic DNA. In the wild-type cells, an expected band was detected with either EcoRI (2.4 kb) or EcoRV (8.0 kb) digestion. In the two TGPIKO strains (IKOA and IKOB), the hybridized band shifted to 4.0 kb (EcoRI) or 4.2 kb (EcoRV), indicating the insertion of the neo cassette into the TGPI gene locus. No bands with the size expected for the wild-type TGPI gene was detected in the TGPIKO cells, suggesting that TGPI gene has been completely disrupted. Similarly, for the analysis of TGP3 gene knockout (Fig. 2B), a TGP3 gene-specific probe was used to hybridize EcoRI or EcoRV digested genomic DNA. Hybridized band was shifted from ~6 kb in wild-type cells to ~7 kb in the TGP3 KO cells (EcoRI digestion), or from ~12 kb to ~7 kb (EcoRV digested), indicating that the neo cassette has been inserted into the TGP3 gene locus. No wild-type band was detected in TGP3KO cells, suggesting a complete disruption of the TGP3 gene. In summary, southern analysis demonstrated that each of the TGP macronuclear genes has been completely disrupted in its corresponding KO cells.

To further confirm the complete knockout of the TGP genes, the KO cells were cultured in medium containing no paromomycin. The rational here is as follows. If the gene has not been completely disrupted (meaning wild-type copy still exists in the macronuclear genome), culturing in no-drug selection medium will result in a reversion of wild-type genes (through the phenotypic assortment). If the gene has been completely disrupted, no such reversion will occur. After growing in no-paromomycin media for two weeks, the KO cells were checked with southern analysis. No reversion of wild-type gene was observed in either TGPIKO or TGP3KO cells (data not shown) confirming that the genes have been completely disrupted in the macronucleus.

G-DNA binding activities lost in the KO cells

*Tetrahymena* macronucleus is transcriptionally active and responsible for all protein expression in the cell. Therefore, complete gene disruption in this nucleus should result in elimination of protein expression of a particular gene. To determine whether the TGP proteins had been eliminated in the KO cells, we checked the G-DNA binding activities of TGP proteins by performing mobility shift assays. Total proteins were extracted from wild-type, TGPIKO, and TGP3KO cells, and tested with mobility shift assay. The results are shown in Figure 3. While the extract from the wild-type cells contained three G-DNA binding activities (TGPI-3), TGPIKO cells exhibited only two G-DNA binding activities (TGP2 and TGP3) but no TGPI activity, indicating that TGPI activity has been abolished in the TGPIKO cells. Similarly in TGP3KO cells, the TGP3 activity disappeared. In summary, the mobility shift assay demonstrated that complete disruption of TGP genes in the macronucleus resulted in the abolishment of the corresponding G-DNA binding activity.
Figure 3. Mobility shift assay of total protein extracts from wild-type (WT), TGP1KO and TGP3KO cells. See the result section for details.
Increased TGP3 activity in TGP1KO cells

In addition to the loss of the particular G-DNA binding activity, the mobility shift assays also showed that there is change in the other G-DNA binding activity in the KO cells (Fig. 3). For example, in TGP1KO cells where TGP1 activity is gone, there is an obvious increase in TGP3 activity compared with wild-type cells. This is very intriguing since TGP1 and TGP3 share significant homology with each other, therefore one may expect that there are some level of redundancy between these two. However, in TGP3KO cells, TGP1 activity does not increase as substantially (as TGP3 does in TGP1KO cells). Instead, the TGP1 activity (position I in wild-type cells) shifts to a lower position (position II), suggesting that the loss of TGP3 may somehow affect the binding of position I complex. One possible interpretation is as follows. In wild-type cells, TGP1 and TGP3 form a heterodimer which binds to G-DNA resulting in position I complex. In the absence of TGP3 (the TGP1/TGP3 heterodimer is gone), TGP1 alone binds to G-DNA to form the lower position II complex.

Neither of the TGP genes is essential for cell growth

Growth rates of the KO cells were determined. Wild type and KO cells were grown in parallel under optimal growth conditions (30°C, 125 rpm shaking, 2% PPYS media). Cell numbers were usually counted every 2 hours during the culturing, and plotted into growth curves shown in Figure 4. Cell doubling rates were determined by curve fitting. The doubling rates of TGP1KO and TGP3KO cells are about 2.75 and 2.6 hours respectively. Both growth rates are slightly higher than that of the wild-type cells (2.45 hours). These data indicate that, while there might be some subtle growth changes in the KO cells, disruption of either TGP1 or TGP3 gene has no profound effect on cell growth. In addition to cell growth rate, we also examined the cell morphology and swimming rate (data not shown), and found no apparent defects in these aspects in the KO cells. Therefore, we conclude that neither of the TGP genes is essential for the vegetative cell growth, and their absence has little effect on cell growth and other aspects of cellular behavior.

Telomeres grow normally in TGP1KO cells

Since G-DNA has been suspected to play a role at telomeres, we studied the telomere growth in the KO cells. *Tetrahymena* telomeres exhibit length dynamics and have an average length of about 1 kb. If *Tetrahymena* cells are cultured continuously in log-phase at 30°C, their telomeres can grow from 1 kb to about 1.6 kb. To determine if telomere growth was affected in the KO cells, we examined the telomere length in the KO cells. KO cells were grown continuously in media containing no paromomycin at 30°C with wild-type cells growing in parallel as the control. Genomic DNA was extracted from the cells periodically, digested and probed with the pTrel plasmid DNA. Since pTrel contains cloned sequences of *Tetrahymena* rDNA gene and
Figure 4. Growth curves of wild-type, TGPIKO and TGP3KO cells. Cells were cultured at 30°C in 2% PPYS medium with constant shaking (125 rpm). Cell numbers were counted at different time points and plotted on a log scale. KO cells have slightly longer doubling times (TGPl: 2.75 hours, TGP3KO: 2.5 hours, vs. wild-type: 2.45 hours).
telomeres, it hybridizes to both the rDNA gene (which will serve as a constant size control) and telomeres. The result of telomere length assay is shown in Figure 5. In the wild-type cells, telomeres grew from 1.0 kb to 1.6 kb in one month period as expected. In TGP1KO cells, after one month, telomeres length also increased from ~1.0 kb to ~1.6 kb similarly as that in wild-type cells. This result indicates that TGP1 gene disruption has little effect on telomere growth. Since we just recently obtained TGP3KO cells, we have not collected enough samples to assay telomere growth in these cells.

Multiple micronuclei in KO cells

TGP proteins are specific DNA binding proteins, and likely localize (as was shown by the nuclear fractionation experiment described in the following result section) and play a role in the nuclei. To determine if the disruption of TGP genes affects the nuclear structure, we examined the nuclear structure by nuclear staining. Nuclei of TGP1KO, TGP3KO and wild-type cells were stained with Hoechst DNA dye and imaged using fluorescence microscopy. Normal Tetrahymena nuclear behavior is summarized in Figure 6A. Initially, the cell contains one micronucleus and one macronucleus. Micronucleus begins to divide to form two micronuclei, followed by macronuclear elongation and separation. The divided nuclei are then distributed into daughter cells. Therefore, in normal cells, there should be no more than two micronuclei. However, in both TGP1KO and TGP3KO cells, we observed large numbers of cells containing more than two micronuclei (sometimes up to 6 as we observed). To obtain statistically significant data, large number (~3000 for each group) of nuclei from each group of cells (wild-type, TGP1KO and TGP3KO) were imaged and counted according to the number of micronuclei in the cell. The results of the micronuclear counting analysis are shown in Figure 6B. In both TGP1KO and TGP3KO cells, the percentages of multiple micronuclei is much higher than those in the wild-type cells. For example, cells containing three micronuclei account for 2.5% in TGP1KO cells, 3.5% in TGP3KO cells, while only less than 1% in the wild-type cells. In the case of cells containing four or more that 4 micronuclei, the percentages of such cells in TGP1KO and TGP3KO cells are also higher than that in wild-type cells. In addition, the number of cells containing two nuclei are also higher in KO cells than that in the wild-type cells. In summary, our data revealed an increased occurrence of multiple micronuclei in both TGP1 and TGP3KO cells. A plausible explanation for this phenotype is that, the micronucleus in the KO cells divides prematurely and faster (may be more than one round before cellular division). In the discussion section, we propose a model to explain how TGP gene disruption may cause such premature and faster-than-usual micronuclear division.
Figure 5. Telomere length assay shows a normal telomere growth in TGPIKO cells. Genomic DNA from continuously log-phase grown cells was extracted, digested with PstI, and probed with pTre1 plasmid DNA which will hybridize to telomeres and internal rDNA.
Figure 6. More aberrant micronuclei in both TGP1KO and TGP3KO cells. (A) Normal and aberrant nuclear structures. Tetrahymena nuclei were stained with Hoechst DNA dye. Images of stained nuclei were captured and processed using the program NIH Image. (B) Percentages of multiple micronuclei in wild-type and KO cells. Cells were nuclear stained and counted according the micronuclei number. The experiment was repeated three times over a one month period. About 1,200 cells were counted each time for each group of the cells. Error bars indicate variations among the three counting. Both TGP1 and TGP3KO cells tend to contain significant more (>two fold) extra (>2) micronuclei (small fluorescent spheres).
TGPl and TGP3 localize mainly to the nuclei

Several approaches have been taken to determine the localization of the TGP proteins in vivo. The first approach was immunostaining using polyclonal antibodies generated against TGP peptides. TGPl peptide antibodies generated this way were found to be not suitable for immunostaining analysis since the antibodies recognize not only TGPl protein but also a major nonspecific antigen (Western blot analysis, data not shown). The second approach we took was to use a newly developed Tetrahymena green fluorescence protein (GFP) fusion expression system (from Dr. Meng-Chao Yao at Fred Hutchinson Cancer Research Center, Seattle, Washington). Vectors containing either TGPl or TGP3 cDNA fused in frame after GFP were constructed and transformed into mating Tetrahymena cells. Both GFPTGPl and GFPTGP3 fusion proteins were found to be expressed in the cytosol. However, neither of the fusion proteins exhibits the corresponding G-DNA binding activity in mobility shift assay, indicating that the fusion proteins are non-functional. Thus, the cytosol localization of the TGP proteins determined by the GFP fusion protein approach is considered to be invalid. We suspect that the nonfunctionality of the GFP fusion proteins could be due to the relatively large size of GFP (26 kD). Therefore, we took a third approach in which we replaced GFP with a 6XHis tag (much smaller than GFP) in the expression vector. This approach is in progress.

We also took a relatively simple approach to determine the in vivo localization of TGP proteins. Using a standard protocol, we fractionated Tetrahymena total protein extract into cytoplasmic and nuclear fractions. The fractions were then tested for TGP activities by mobility shift assay. The results are shown in Figure 7. The cytoplasmic fraction contains TGP2 activity (which is a dihydrolipoamide dehydrogenase [DLDH] localized mainly in mitochondrial membrane, thus is expected to be in the cytoplasmic fraction), but almost no TGPl and TGP3 activities. In contrast, the nuclear fractions contain TGPl and TGP3 activities but no TGP2 activity. These data demonstrate that TGPl and TGP3 activities localize mainly in the nuclei.

DISCUSSION

In this paper we investigated the biological roles of two novel G-DNA binding proteins (TGPl and TGP3) from Tetrahymena thermophila. Each of the TGP genes has been completely disrupted in the somatic nucleus. KO cells grow at near normal rate indicating that neither of the genes is essential for cell survival. However, nuclear staining analysis revealed an increased occurrence of multiple micronuclei, suggesting a faulty control of the micronuclear division in the KO cells. In addition to knockout analyses, we determined the nuclear localization of TGP proteins.
Figure 7. TGP1 and TGP3 activities localize mainly to the nuclei. *Tetrahymena* cells were fractionated into cytoplasmic and nuclear fractions according to a standard method. Fractions were lysed and tested by mobility shift assay. Majority of TGP1 and TGP3 activities was found in the nuclear fractions, whereas TGP2 which is a dihydrolipoamide dehydrogenase was found in the cytoplasm. C: cytoplasmic fraction; N1: nuclear fraction (first lysis); N2: nuclear fraction (second lysis).
Figure 8. A model for TGP function in micronuclear division. Telomeres of two sister chromatids are held together through formation of G-DNA structure. TGP protein could recognize and bind to the G-DNA strengthening the association between the telomeres. The TGP/G-DNA complex may serve as a regulatory role in the sister chromatids segregation during mitosis. Disruption of the TGP protein will weaken the binding force between the telomeres and thus cause problems in the chromosome segregation.

Figure 9. A model for TGP function in macronuclear minichromosome organization. Telomeres of the minichromosomes can form G-DNA structure which is then recognized and bound by TGP s. TGP proteins could associate with the nuclear matrix, therefore, anchoring the minichromosomes to the matrix. Thus, instead of floating around in the nucleus, hundreds of thousands of minichromosomes get organized. This organization may be important for the transcription of genes on the minichromosomes.
TGPs in micronuclear division

Like nuclei in other organisms, the *Tetrahymena* micronucleus is diploid and divide mitotically distributing chromosomes equally into daughter cells. Mitotic chromosome segregation is known to be well regulated (15, 19). Recently, telomere-telomere association is suggested to be involved in this process (11, 12). A study by Kirk et al. (18) showed that a mutation in *Tetrahymena* telomeric DNA repeats (T$_2$G$_4$ changed to T$_4$G$_4$) caused severely delayed micronuclear division. Cytological analysis of the nuclei revealed that sister chromatids adhere to each other at telomeres until late anaphase, and became stretched up to twice the normal length when finally separated, indicating stronger physical association between the mutant chromatids at telomeres. This study established a potential role for telomere-telomere association in mitotic chromosome segregation.

However, little is known about how telomeres associate with each other *in vivo*. One model (Figure 8) would be that guanine-rich strands or 3' G-overhangs of telomeres are bonded together by formation of G-DNA between the sequences. Once the G-DNA is formed between the telomeres, protein components (especially those G-DNA binding proteins such as TGPl and TGPS) may be recruited to the site to stabilize the association between the telomeres. This G-DNA/protein complex will be intact holding the chromosomes together during much of the mitosis until anaphase when the complex finally dissemble and the chromosomes separate. In this way, the complex serves a regulatory role in mitotic chromosome segregation. This model explains the delayed micronuclear division phenotype observed in the telomeric DNA mutant, since the mutation of telomere sequence (from T$_2$G$_4$ to T$_4$G$_4$) may somehow strength the G-DNA-mediated telomere-telomere association between the two sister chromatids and thus cause problems in chromosome segregation.

This model also nicely explains the multiple micronuclei phenotype we observed in the KO (either TGPl or TGPS) cells. We envision the following scenario in TGPKO cells. In the absence of the TGP protein, the G-DNA structure formed between the telomeres of chromatids will become susceptible to cellular enzymes such as helicases and nucleases resulting in a weakened association between the telomeres. This weakened association will thus cause the sister chromatids to separate prematurely (well before anaphase). The prematurely separated micronuclei may initiate another round of DNA synthesis and divide one more time before the cellular division resulting in the multiple micronuclear phenotype. Thus, according to this model, TGP proteins play roles in mitotic chromosome division through binding to the G-DNA formed between telomeres of sister chromatids.

TGPs in macronuclear organization

TGP proteins are relatively abundant and it is likely that these proteins localize in both micro- and macronucleus. Although the nuclei staining analysis does not reveal obvious changes
in macronuclear size or structure, this does not exclude a role of TGP proteins in macronucleus. Unlike the micronucleus, the macronucleus divides amitotically distributing chromosomes randomly to the daughter cells. Hence, there is probably no need to tightly regulate the chromosome separation in the macronucleus. So what is the role of the TGP proteins in the macronucleus? It should be noted that, in addition to its amitotic division, the macronucleus has another interesting characteristic: its genome is fragmented and greatly amplified. The average length of a macronuclear chromosome is about 600 kB, 20 times smaller than that of an average micronuclear chromosome, and the average gene copy number is about 45 (26). Thus, in the macronuclear genome, there are at least 10,000 minichromosomes. These chromosomes are responsible for synthesizing the proteins that cause the cell phenotype. This enormous number of minichromosomes has to be organized to function in a spatially and temporally conducted fashion. One common feature of these minichromosomes is that they all contain telomeres. We speculate the TGP protein may play a role in minichromosome organization. We envision the following scenario. Telomeres of the minichromosomes can form G-DNA structure which is then recognized and bound by TGPs. TGP proteins could associate with the nuclear matrix, therefore, anchoring the minichromosomes to the matrix. Thus, instead of floating around in the nucleus, hundreds of thousands of minichromosomes get organized. Although currently we do not have solid experimental data to validate this model, there is some interesting information supporting it. First, both TGP1 and TGP3 contain sequences capable of forming coil-coiled structure. TGP proteins may interact with nuclear matrix proteins through this coil-coiled motif. Second, TGP proteins may be themselves nuclear matrix proteins, since TGP1 was found to have a weak similarity with yeast NUF1, a possible nuclear matrix protein (24, 25).

Redundancy and interaction between TGP1 and TGP3?

The homology between TGP1 and TGP3 suggests that the proteins may have similar functions in vivo, and furthermore there could be certain level of functional redundancy between these two proteins. Indeed, we found that TGP3 activity increases significantly in TGP1 knockout cells, suggesting that TGP3 may compensate for the loss of TGP1 by increasing its expression. In TGP3KO cells where TGP3 activity is gone, the TGP1 activity also incise slightly. The possible functional redundancy between TGP1 and TGP3 may explain the relatively mild phenotype we observed in the KO cells. It will be of high interest to disrupt both genes in the somatic nucleus and look for more severe phenotypes. Such double gene disruption experiment is underway.

In addition to the potential redundancy between the two proteins, TGP1 and TGP3 may interact with each other in vivo. In TGP3 KO cells, TGP1 activity shifts to a lower position in mobility shift assay gels (Figure 3). We interpret this result as follows. In wild-type cells, TGP1 and TGP3 interact to form a heterodimer which can bind to G-DNA resulting in the position I complex in the mobility shift assay gel. When TGP3 is knocked out (TGP3KO), this heterodimer
is gone resulting in a lower (position II) shifted complex (TGP1/G-DNA). If TGP1 is knocked out (TGP1KO), both the position I and II complex will be gone with only TGP3/G-DNA complex left. Thus the shifted band in TGP3KO cells suggests a potential interaction between TGP1 and TGP3. Two other lines of evidence support the interaction between the proteins. First, TGP1 and TGP3 copurify during a three-column purification process. Second, the position I complex excised from a mobility shift assay gel contains two protein bands (TGP1 and TGP3) (data not shown).

REFERENCES


GENERAL CONCLUSIONS

G-DNA is a novel four-stranded DNA structure containing motifs known as the G-quartet. Even though its in vivo presence remains to be determined, G-DNA is proposed to play roles in many diverse biological processes including telomere function, recombination and gene regulation. I have been studying two novel specific G-DNA-binding proteins.

Identification, purification and cloning of TGP

Chapter 1 described the purification, partial biochemical characterization and molecular cloning of TGPI, which was originally identified by Schierer and Henderson as a parallel G-DNA specific binding protein (68). TGPI was partially purified by three chromatographies, including a G-DNA affinity column. Biochemical characterization of partially purified protein confirms that TGPI is a specific G-DNA-binding protein. The cDNA coding for the protein was cloned based on a peptide sequence obtained from direct internal peptide sequencing. The cDNA encodes a novel 726 amino acid long protein which has no significant homologs in available protein databases. During TGPI purification, an additional G-DNA binding activity referred to as TGP3 was found to copurify with TGPI. The molecular cloning of TGP3 cDNA is described in Chapter 2. The TGP3 cDNA encodes a 365 amino acid long proteins. Sequence comparison showed that TGP3 shares significant homology (34% identity and 44% similarity) with TGPI but not any other proteins in the available databases. To our knowledge, this represents the first example of homology among known G-DNA binding proteins.

More intriguingly, detailed sequence analysis revealed that both proteins share a sequence pattern containing an extensively hydrophilic and basic region flanked by two repetitive and homologous motifs. This sequence pattern is not found in any other proteins, suggesting that this may represent a novel protein motif. Since this domain covers most of the TGP3 sequence, it is likely that this domain is responsible for the G-DNA binding activity. However, it is currently unclear if this whole sequence pattern is required for the binding. It is possible that one of the repetitive and homologous motifs is sufficient for the binding. To identify the minimal G-DNA binding motif, serial deletion analysis should be performed.

The novel structural motif found in TGPI and TGP3 offers an excellent model to study the structural basis of G-DNA/protein interaction or binding. The minimal domain identified by deletion analysis can be engineered (to remove stop codons) and expressed in E. coli to obtain large amount of the protein. The expressed motif can be crystallized with the G-DNA and used for high-resolution structural analyses (NMR or X-ray crystallography). These structural studies will provide invaluable information about the structure of the protein motif and, more importantly, how it recognizes and specifically binds to the G-DNA structure.
It is somewhat disappointing not to find TGP homologs in other organisms including the yeast and \textit{C. elegans} whose complete genomes have been sequenced. There are two possible explanations for this no-homolog situation. One explanation is that the TGP proteins (especially the shared motif) may be specific for \textit{Tetrahymena}. It is reasonable to assume that \textit{Tetrahymena} uses specific proteins to maintain some of its characteristics such as the nuclear dualism. TGP\textsubscript{s} may be among such proteins. If this is the case, it would be more feasible to look for TGP homologies in other ciliates (such \textit{Oxytricha} and \textit{Euplotes}) which share these characteristics with \textit{Tetrahymena}. Such cloning work is already in progress in the lab. The other explanation for why no TGP homologs were found is that proteins from other organisms may be structurally (but not in primary sequence) similar to TGP\textsubscript{s}. This has been the case for the \textit{Oxytricha} telomere-end binding protein which has no apparent sequence homologs but contains structural motifs called oligonucleotide/oligosaccharide folds shared by many other proteins (69). Identifying structural homologs of TGP proteins may require the elucidation of TGP structures first.

\textbf{Biological roles of TGP proteins}

In chapter 3, we studied the biological functions of TGP\textsubscript{1} and TGP\textsubscript{3}. Using a newly developed \textit{Tetrahymen} gene knockout technique, we have generated macronuclear gene knockout (KO) strains separately for each of the two genes. Southern analysis and mobility shift assay showed that each of the genes has been completely disrupted in its corresponding KO strains (TGP1KO and TGP3KO). The KO cells grow at near wild-type rate indicating that neither of the genes is essential for cell growth. However, nuclear staining analysis revealed that the KO cells have an increased occurrence (more than two fold) of extra micronuclei, suggesting a faulty control of the micronuclear division in the KO cells.

How might TGP gene disruption cause the multiple micronuclear phenotype? We propose a model for functions of TGPs in the micronuclear division. In this model, telomeres of two sister chromatids are brought together by formation of a dimeric G-DNA. Once the G-DNA is formed, protein components (especially those G-DNA binding proteins such as TGP\textsubscript{1} and TGP\textsubscript{3}) may be recruited to the site to stabilize the association between the telomeres. This G-DNA/protein complex will be intact holding the chromosomes together during much of the mitosis until anaphase when the complex finally disassemble and the chromosomes separate. In this way, the complex serves a regulatory role in mitotic chromosome segregation. Disruption of TGP will result in a weaker telomere-telomere association, which may cause premature separation of the sister chromatids. Although attractive and able to explain the phenotypes we observed in TGP KO cells, this model needs to be rigorously tested. For example, it will be very important to determine if TGPs localize at the micronuclear telomeres. This may be done by performing FISH (fluorescent \textit{in situ} hybridization) on the micronuclear chromosomes using TGP antibodies.
Mobility shift assay of extracts from KO cells show that TGP3 activity increases significantly in TGPI KO cells, and similarly, in TGP3 KO cells where TGP3 activity is gone, TGPI activity also incises slightly. These data suggest that TGPI (or TGP3) may compensate for the loss of the other protein by increasing its own expression. Such functional compatibility between TGPI and TGP3 is consistence with the fact that the two proteins are highly homologous to each other. This may also explain the relatively mild phenotypes in the KO cells. To further investigate the biological roles of the proteins, it will be of high interest to perform a double knockout of these two genes. We have finished constructing all the necessary knockout vectors for the double gene KO experiment. We are now working to transform these vectors into the *Tetrahymena* germ-line nucleus.

We have also determined the localization of TGP proteins in vivo. Nuclear fractionation assay clearly shows that TGPI and TGP3 localize mainly (if not exclusively) in the nuclear fraction. This is consistent with the G-DNA binding nature of the proteins. However, this assay does not tell us whether the proteins localize in one of the nuclei or in both. Currently we are using a His-tagging expression system to further study the protein localization. This study may also provide large amount of His-tagged TGPs which can be used to raise antibodies.

REFERENCES