Structure, Vol. 12, 1877–1879, October, 2004, ©2004 Elsevier Ltd. All rights reserved. DOI 10.1016/j.str.2004.07.015

Prediction of Multiple Tandem OB-Fold Domains in Telomere End-Binding Proteins Pot1 and Cdc13

Douglas L. Theobald and Deborah S. Wuttke*
Department of Chemistry and Biochemistry
University of Colorado at Boulder
Boulder, Colorado 80309

Summary

The heterodimeric *Oxytricha nova* telomere end binding protein, the original telomere end binding protein characterized, contains four OB-fold domains used for recognition of single-stranded telomeric DNA. In contrast, only solitary OB-fold domains have been found in the telomere end binding proteins from yeast and higher eukaryotes. Using a sliding-window algorithm coupled with sequence profile-profile analysis, we provide support for the existence of multiple OB-fold domains in two other telomeric ssDNA binding proteins, vertebrate Pot1 and budding yeast Cdc13. This common usage of multiple, tandem OB-fold domains in telomeric end binding proteins extends the known evolutionary conservation of eukaryotic end-protection mechanisms.

Introduction

Several eukaryotic proteins recognize the single-stranded DNA telomeric overhang of chromosomes, including the ciliate telomere end binding proteins (TEBP), the metazoan Pot1 proteins, and budding yeast Cdc13 proteins (Baumann and Cech, 2001; Froelich-Ammon et al., 1998; Gottschling and Zakian, 1986; Nugent et al., 1996; Wei and Price, 2004). These proteins bind the telomeric overhang with both high specificity and strong affinity (Anderson et al., 2003; Classen et al., 2003; Lei et al., 2002) and are critical for chromosome capping, end protection, and telomere-length regulation (Cervantes and Lundblad, 2002; Loayza and de Lange, 2003; Smogorzewska and de Lange, 2004). Each of these proteins uses an OB-fold domain, one of the most common protein domains, to recognize telomeric, G-rich ssDNA (Horvath et al., 1998; Lei et al., 2003; Mitton-Fry et al., 2004).

Telomeric OB-fold domains participate in both the binding of ssDNA and the regulation of other telomeric factors. In TEBP, four OB-fold domains in two subunits either bind ssDNA directly or act as protein-protein interaction modules (Horvath et al., 1998). The Pot1 proteins were originally identified from weak sequence similarity between their N termini and the N-terminal OB-fold domain of the TEBP α subunit (Baumann and Cech, 2001). The Cdc13 DNA binding domain also shares weak sequence similarity to the N-terminal Pot1 OB-fold domain, as determined from sequence profile analysis (Theobald et al., 2003a). Unlike TEBP, only one OB-fold domain has been found to date in both Cdc13 and Pot1 (Lei et al., 2003; Mitton-Fry et al., 2002), although the possibility of a second Pot1 OB-fold domain has been postulated (Wei and Price, 2004).

As exemplified by TEBP and other ssDNA binding proteins, such as *E. coli* SSB and human RPA, multiple OB-fold domains are common in nucleic acid recognition. However, OB-fold domains are notoriously difficult to detect based upon sequence similarity alone, and most proteins containing this structural motif share little sequence similarity (Murzin, 1993; Theobald et al., 2003b). Nevertheless, due to evolution's opportunistic nature of reusing modified domains in interacting networks (Doolittle, 1995), we expected the possibility of finding cryptic OB-fold domains in other telomere-associated proteins.

Sequence-profile methods benefit from the additional information conferred from a collection of related sequences, and profile-based sequence analyses significantly outperform pairwise methods in detecting remote homology (Park et al., 1998). A recently developed profile-profile comparison program, COMPASS (used in this work), is currently one of the most sensitive methods for remote homolog detection (Sadreyev and Grishin, 2003). However, multiple tandem copies of a domain provide a difficult challenge for sequence comparison. Many current profile programs, such as COMPASS, report only optimal matches. Thus, additional weaker-scoring yet significant regions, representing multiple appearances of a domain, may go undetected.

Using sequence-profile analyses directed at tandem copies of domains, we provide evidence here that both the vertebrate Pot1 proteins and Cdc13 proteins likely contain additional multiple OB-fold domains.

Results and Discussion

We developed a straightforward "sliding-window" algorithm, coupled with the COMPASS program, to determine if multiple regions of a longer sequence alignment match a profile of a smaller domain. Using this technique, full-length Cdc13 and Pot1 proteins were scanned with smaller OB-fold domains using the COMPASS program (Figure 1). As anticipated, our profile-based analysis gave peaks corresponding within 15 residues to the known OB-fold domains in the N terminus of the full-length vertebrate Pot1 proteins (Figure 1A and 1B) and in the C terminus of full-length yeast Cdc13 (Figure 1C).

Furthermore, our analysis of Pot1 revealed two additional OB-fold domains, one strongly and one weakly predicted, C-terminal to the Pot1 OB-fold domain previously identified. The peaks corresponding to the Pot1 OB-fold domains were found independently by scanning with both the Cdc13 OB-fold domain (Figure 1A) and the bacterial SSB OB-fold domain (Figure 1B), observations that strengthen the confidence assigned to these predictions. Whether there are in fact multiple OB-fold do-

*Correspondence: deborah.wuttke@colorado.edu

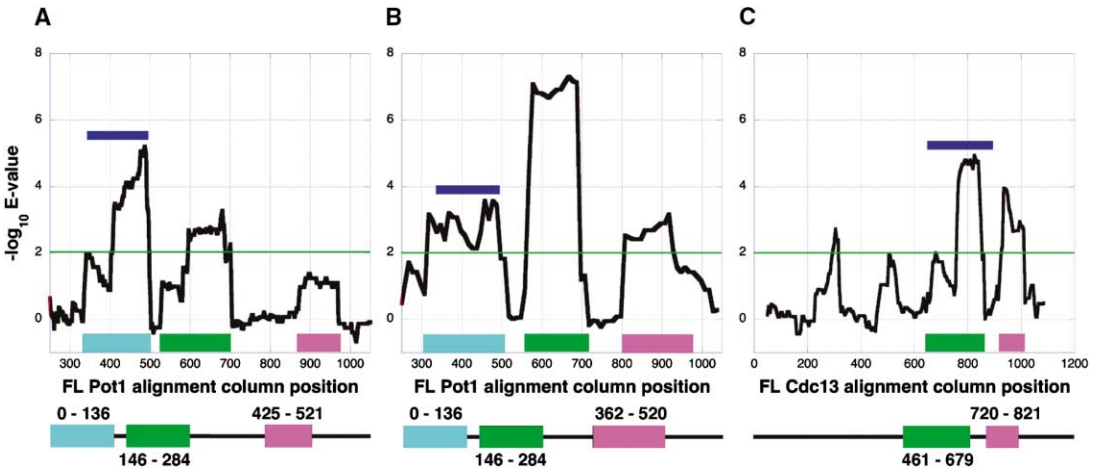

Figure 1. OB-Fold Domain Predictions

The negative $\log_{10}$ of the E-value plotted against window position in full-length alignments of Pot1 and Cdc13. E-values < 0.01 are indicated by values above the horizontal green line.

(A) Full-length Pot1 (FL) alignment scanned with the Cdc13 OB-fold profile. The first of the three peaks, indicated with a blue bar, corresponds to the OB-fold domain previously identified in Pot1 (residues 1–174 in *Homo sapiens* Pot1) (Lei et al., 2003).

(B) Full-length Pot1 alignment scanned with the bacterial SSB OB-fold domain.

(C) Full-length Cdc13 alignment scanned with the Pot1 OB-fold domain. The first large peak, indicated with a blue bar, corresponds to the known OB-fold ssDNA binding domain in Cdc13 (residues 501–672 in *S. cerevisiae* Cdc13) (Mitton-Fry et al., 2002).

(A–C) Below each plot are schematics showing the approximate predicted OB-fold domain boundaries for the corresponding analyses above. Cyan, green, and magenta boxes correspond to major peaks in the MarkovPolo plots, where they are colored similarly. The human Pot1 protein (for 1a and 1b) and the *S. cerevisiae* Cdc13 protein (for 1c) are used as reference sequences for residue numbers.

mains in the human Pot1 protein should be known soon, as the crystal structure has recently been solved (M. Lei, E. Podell, and T. Cech, personal communication). Additionally, we have identified a putative OB-fold domain at the C terminus of Cdc13, C-terminal to the known OB-fold DNA binding domain (Figure 1B). As negative controls, we have been unable to detect false positives after scanning these OB-fold alignments and the full-length proteins against a database of alignments, including all known RNA recognition motif domains (RRMs), KH-domains, dsRNA binding domains (RBDs), SH3-domains, and PDZ domains.

The presence of three tandem OB-fold domains in Pot1, closely analogous to the ciliate TEBP α subunit, suggests that these domains are involved in yet uncharacterized telomeric functions. For example, the second OB-fold domain of TEBP is necessary for 3′-end recognition of telomeric ssDNA (Horvath et al., 1998). The third TEBP α OB-fold domain, however, acts as a protein-protein interaction module linking to the ssDNA binding OB-fold domain of the TEBP β subunit. In turn, the OB-fold domain of the TEBP β subunit is known to modulate access by telomerase to the end of the chromosome in vitro and likely is involved in telomere length regulation (Fang and Cech, 1993; Froelich-Ammon et al., 1998). In another structural context, the third OB-fold domain of TEBP α acts as a homodimerization domain in a complex that binds telomeric ssDNA very differently from the heterodimeric α-β-ssDNA complex (Peersen et al., 2002). Intriguingly, a telomeric factor, PIP1/PTOP, which binds to the C terminus of Pot1 and recruits it to the telomere, has been identified recently in human cells (Liu et al., 2004; Ye et al., 2004). Our prediction of a third Pot1 OB-fold domain, localized to residues 362–521 in the human protein (Figure 1A and 1B), corresponds well to the genetically determined PIP1-interaction domain of Pot1, which begins between residues 312 and 428 and ends at or before the C-terminal residue 634 (Ye et al., 2004). Our profile-based analysis further extends the known evolutionary conservation of telomere-associated proteins and indicate that additional OB-fold domains in telomere end binding proteins may be important for telomere capping and length regulation.

Experimental Procedures

Telomeric Sequence Profiles

CLUSTAL alignments (Chenna et al., 2003) were made of close homologs of the known OB-fold domains of budding yeast Cdc13 (residues 497–694), vertebrate Pot1 (residues 7–169 of the human protein), and the *E. coli* single-stranded DNA binding protein (SSB, residues 2–145), as well as alignments of full-length Cdc13 and full-length Pot1, as previously described (Theobald et al., 2003a). Close fungal homologs of *Saccharomyces cerevisiae* Cdc13 were identified via BLAST searches of five yeast genomes (Cliften and Johnston, 2003). The Pot1 alignments include eight vertebrate sequences, including human, *Macaca*, mouse, rat, and chicken. The *E. coli* OB-fold alignment includes over 200 close bacterial homologs determined from BLAST searches of the NCBI nonredundant protein database. Sequence alignments are available from the authors upon request.

Algorithm for Detecting Tandem Domains

Multiple copies of a domain within a longer protein are identified by scanning and iteratively scoring an alignment of a full-length protein with a smaller domain-sized alignment, using a sequence-profile program such as COMPASS (Sadreyev and Grishin, 2003). The resulting E-values, rescaled for total alignment length, are plotted against the window position in the full-length alignment. When the window position coincides with a domain homologous to the smaller query domain, a peak is observed in the plot of the scores. It has been shown that the COMPASS E-values correspond well with

empirically determined false positive rates (Sadreyev and Grishin, 2003), and our analysis should have a similar false positive rate after correction for the total length of the scanned profile. This widely applicable, intuitive method provides a visual representation that allows for fast identification of multiple domain repeats (Figure 1).

We have implemented this technique for general use in a program called MarkovPolo. MarkovPolo applies this algorithm as a wrapper application using several publicly available packages for constructing and scoring the profiles. Currently supported profile packages include COMPASS, SAM (Karplus et al., 1998; Park et al., 1998), HMMER (Eddy, 1998), PSI-BLAST, and prof_sim (Yona and Levitt, 2002). MarkovPolo is available online via a web-server and for local use as a command-line utility at http://monkshood.colorado.edu/markovpolo/.

Acknowledgments

We would like to thank Ming Lei and Thomas Cech for helpful discussions. We gratefully acknowledge the NIH (D.S.W.), the Arnold and Mabel Beckman Foundation (D.S.W.) and the American Cancer Society (D.L.T.) for support.

Received: June 18, 2004
Revised: July 21, 2004
Accepted: July 31, 2004
Published: October 5, 2004

References

Anderson, E.M., Halsey, W.A., and Wuttke, D.S. (2003). Site-directed mutagenesis reveals the thermodynamic requirements for single-stranded DNA recognition by the telomere-binding protein Cdc13. Biochemistry *42*, 3751–3758.

Baumann, P., and Cech, T.R. (2001). Pot1, the putative telomere end-binding protein in fission yeast and humans. Science *292*, 1171–1175.

Cervantes, R.B., and Lundblad, V. (2002). Mechanisms of chromosome-end protection. Curr. Opin. Cell Biol. *14*, 351–356.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., and Thompson, J.D. (2003). Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. *31*, 3497–3500.

Classen, S., Lyons, D., Cech, T.R., and Schultz, S.C. (2003). Sequence-specific and 3′-end selective single-strand DNA binding by the *Oxytricha nova* telomere end binding protein alpha subunit. Biochemistry *42*, 9269–9277.

Clifton, P., and Johnston, M. (2003). Washington University Genome Sequencing Center. (http://www.genome.wustl.edu/projects/yeast/)

Doolittle, R.F. (1995). The multiplicity of domains in proteins. Annu. Rev. Biochem. *64*, 287–314.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics *14*, 755–763.

Fang, G., and Cech, T.R. (1993). *Oxytricha* telomere-binding protein: DNA-dependent dimerization of the alpha and beta subunits. Proc. Natl. Acad. Sci. USA *90*, 6056–6060.

Froelich-Ammon, S.J., Dickinson, B.A., Bevilacqua, J.M., Schultz, S.C., and Cech, T.R. (1998). Modulation of telomerase activity by telomere DNA-binding proteins in *Oxytricha*. Genes Dev. *12*, 1504–1514.

Gottschling, D.E., and Zakian, V.A. (1986). Telomere proteins: specific recognition and protection of the natural termini of *Oxytricha* macronuclear DNA. Cell *47*, 195–205.

Horvath, M.P., Schweiker, V.L., Bevilacqua, J.M., Ruggles, J.A., and Schultz, S.C. (1998). Crystal structure of the *Oxytricha nova* telomere end binding protein complexed with single strand DNA. Cell *95*, 963–974.

Karplus, K., Barrett, C., and Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. Bioinformatics *14*, 846–856.

Lei, M., Baumann, P., and Cech, T.R. (2002). Cooperative binding of single-stranded telomeric DNA by the Pot1 protein of *Schizosaccharomyces pombe*. Biochemistry *41*, 14560–14568.

Lei, M., Podell, E.R., Baumann, P., and Cech, T.R. (2003). DNA self-recognition in the structure of Pot1 bound to telomeric single-stranded DNA. Nature *426*, 198–203.

Liu, D., Safari, A., O'Connor, M.S., Chan, D.W., Laegeler, A., Qin, J., and Songyang, Z. (2004). PTOP interacts with POT1 and regulates its localization to telomeres. Nat. Cell Biol. *6*, 673–680.

Loayza, D., and de Lange, T. (2003). POT1 as a terminal transducer of TRF1 telomere length control. Nature *423*, 1013–1018.

Mitton-Fry, R.M., Anderson, E.M., Hughes, T.R., Lundblad, V., and Wuttke, D.S. (2002). Conserved structure for single-stranded telomeric DNA recognition. Science *296*, 145–147.

Mitton-Fry, R.M., Anderson, E.M., Theobald, D.L., Glustrom, L.W., and Wuttke, D.S. (2004). Structural basis for telomeric single-stranded DNA recognition by yeast Cdc13. J. Mol. Biol. *338*, 241–255.

Murzin, A.G. (1993). OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. EMBO J. *12*, 861–867.

Nugent, C.I., Hughes, T.R., Lue, N.F., and Lundblad, V. (1996). Cdc13p: a single-strand telomeric DNA-binding protein with a dual role in yeast telomere maintenance. Science *274*, 249–252.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., and Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J. Mol. Biol. *284*, 1201–1210.

Peersen, O.B., Ruggles, J.A., and Schultz, S.C. (2002). Dimeric structure of the *Oxytricha nova* telomere end-binding protein alpha-subunit bound to ssDNA. Nat. Struct. Biol. *9*, 182–187.

Sadreyev, R., and Grishin, N. (2003). COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. J. Mol. Biol. *326*, 317–336.

Smogorzewska, A., and de Lange, T. (2004). Regulation of telomerase by telomeric proteins. Annu. Rev. Biochem. *73*, 177–208.

Theobald, D.L., Cervantes, R.B., Lundblad, V., and Wuttke, D.S. (2003a). Homology among telomeric end-protection proteins. Structure *11*, 1049–1050.

Theobald, D.L., Mitton-Fry, R.M., and Wuttke, D.S. (2003b). Nucleic acid recognition by OB-fold proteins. Annu. Rev. Biophys. Biomol. Struct. *32*, 115–133.

Wei, C., and Price, C.M. (2004). Cell cycle localization, dimerization, and binding domain architecture of the telomere protein cPot1. Mol. Cell. Biol. *24*, 2091–2102.

Ye, J.Z.-Y., Hockemeyer, D., Krutchinsky, A.N., Loayza, D., Hooper, S.M., Chait, B.T., and de Lange, T. (2004). POT1-interacting protein PIP1: a telomere length regulator that recruits POT1 to the TIN2/TRF1 complex. Genes Dev. *18*, 1649–1654.

Yona, G., and Levitt, M. (2002). Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. J. Mol. Biol. *315*, 1257–1275.