STRIKING SIMILARITIES IN DIVERSE TELOMERASE PROTEINS REVEALED BY COMBINING STRUCTURE PREDICTION AND MACHINE LEARNING APPROACHES

JAE-HYUNG LEE[1,2†], MICHAEL HAMILTON[5], COLIN GLEESON[6], CORNELIA CARAGEA[3,4], PETER ZABACK[1,2] , JEFFRY D. SANDER[1,2], XUE LI[1], FEIHONG WU[1,3,4], MICHAEL TERRIBILINI[1,2] ,VASANT HONAVAR[1,3,4], DRENA DOBBS[1,2,4]

[1]Bioinformatics & Computational Biology Program, L.H. Baker Center for Bioinformatics & Biological Statistics, [2]Dept. of Genetics, Development & Cell Biology, [3]Dept. of Computer Science, [4]Artificial Intelligence Research Lab & Center for Computational Intelligence, Learning & Discovery, Iowa State University, Ames, IA, 50010, USA

[5]Dept. of Computer Science, Colorado State University, Fort Collins, CO 80523, USA

[6]Dept. of Biological Sciences, Univ. of Illinois, Chicago, IL, 60607, USA

Telomerase is a ribonucleoprotein enzyme that adds telomeric DNA repeat sequences to the ends of linear chromosomes. The enzyme plays pivotal roles in cellular senescence and aging, and because it provides a telomere maintenance mechanism for ~90% of human cancers, it is a promising target for cancer therapy. Despite its importance, a high-resolution structure of the telomerase enzyme has been elusive, although a crystal structure of an N-terminal domain (TEN) of the telomerase reverse transcriptase subunit (TERT) from *Tetrahymena* has been reported. In this study, we used a comparative strategy, in which sequence-based machine learning approaches were integrated with computational structural modeling, to explore the potential conservation of structural and functional features of TERT in phylogenetically diverse species. We generated structural models of the N-terminal domains from human and yeast TERT using a combination of threading and homology modeling with the *Tetrahymena* TEN structure as a template. Comparative analysis of predicted and experimentally verified DNA and RNA binding residues, in the context of these structures, revealed significant similarities in nucleic acid binding surfaces of *Tetrahymena* and human TEN domains. In addition, the combined evidence from machine learning and structural modeling identified several specific amino acids that are likely to play a role in binding DNA or RNA, but for which no experimental evidence is currently available.

1. Introduction

In most eukaryotes, a remarkable ribonucleoprotein enzyme, telomerase, is responsible for the synthesis and maintenance of telomeres, the ends of linear chromosomes [1, 2, 3]. Many exciting discoveries have been made in telomerase biology since 1984, when the enzyme was first identified in the ciliate,

[†] Corresponding author

Tetrahymena thermophila, by Greider and Blackburn [4]. Recently, pivotal roles for telomerase in signaling pathways that regulate cancer, stress response, apoptosis and aging have been demonstrated [5, 6, 7, 8].

Two essential roles of telomeres are protecting or "capping" chromosome ends and facilitating their complete replication (reviewed in 1, 2, 3). Typically, telomeres consist of arrays of simple DNA sequence repeats, ranging from ~50 copies of 5'-TTGGGG-3' in *Tetrahymena*, to ~1000 copies of 5'-TTAGGG-3' in humans and other vertebrates. The sequence of telomeric repeats is specified by an RNA template (TER), which varies in length from ~160 nts in ciliates to ~1500 nts in vertebrates, and is an essential component of the catalytically active form of telomerase [2, 5]. Human telomerase is composed of hTER and two bound proteins, the telomerase reverse transcriptase component (hTERT) and dyskerin [9]. The regulation of telomerase activity involves interactions with a variety of other cellular proteins, many of which are essential for telomere homeostasis [8, 10].

Telomerase is a promising target for cancer therapy because it is generally present in very low levels in normal somatic cells, but it is highly active in many human malignancies [11]. Telomerase targeting strategies have included short interfering RNA (siRNA) knockdown of endogenous hTER and a combination of siRNA and expression of mutant forms of the hTER RNA, which become incorporated into the enzyme and inhibit proliferation in variety of different human cancer cell lines [11].

Despite its obvious clinical importance, currently there are no experimentally determined structures for the telomerase ribonucleoprotein complex or for telomerase complexes bound to telomeric DNA substrates, presumably because these are multisubunit structures. The telomerase reverse transcriptase component, TERT, is generally thought to consist of four functional domains (see Figure 1): the essential N-terminal (TEN) domain, an RNA-binding domain (TRBD), reverse transcriptase (RT), and a C-terminal extension (TEC). Recently, a crystal structure of the essential N-terminal domain of TERT from *Tetrahymena* has been reported [12] and appears to represent a novel protein fold. Several conserved sequence motifs have been identified within the TEN domain on the basis of multiple sequence alignments and mutagenesis experiments [13, 14]. In addition, experiments directed at mapping DNA and RNA binding sites within TERTs from several organisms have identified specific amino acids that appear to contact either the DNA template or the RNA component [reviewed in 3]. In human telomerase, the TEN domain binds both DNA, specifically interacting with telomeric DNA substrates, and RNA, apparently binding in a non-sequence specific manner [12].

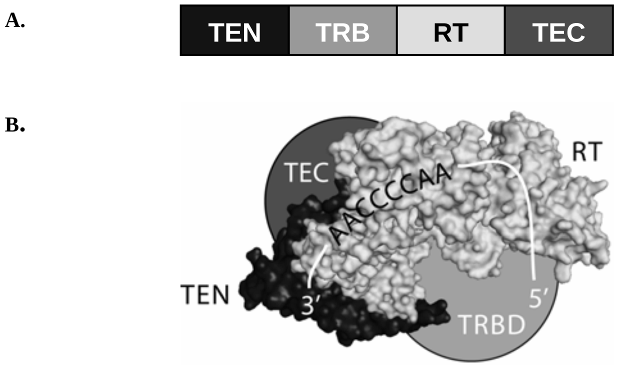


Figure 1. TERT domain architecture. A) The telomerase reverse transcriptase (TERT) comprises 4 functional domains: essential N-terminal (TEN) domain, RNA-binding domain (TRBD), reverse transcriptase (RT), and C-terminal extension (TEC). **B)** Cartoon illustrating TERT domain organization, and the RNA template (TER). The TEN domain is *Tetrahymena* structure (PDB ID: 2B2A), and RT domain is from HIV-RT (PDB ID: 3HVT). Figure modeled after Collins, 2006 [2].

Although vertebrate TEN domain sequences share a high degree of sequence similarity, the TEN domains from more diverse species share very little sequence similarity (<30% identity), suggesting that a homology modeling approach to predicting the structure of the human TEN domain would be difficult. However, an alignment of the N-terminal sequences of TERTs from organisms ranging from human to *T. thermophila* to *S. cerevisiae*, revealed several highly conserved residues distributed throughout the N-terminal domain, suggesting that TEN domains from diverse organisms may share similar architectures [12]. Based on this suggestion, we set out to test the hypothesis that the N-terminal domains of TERTs in diverse organisms not only share a similar overall three-dimensional fold, but may also have phylogenetically conserved DNA and RNA binding surfaces. We used a strategy in which comparative protein structural modeling approaches were integrated with sequence-based machine learning approaches for predicting DNA or RNA binding residues.

2. Datasets, Materials and Methods

2.1 Datasets

RNA-protein interface dataset

A dataset of protein–RNA interfaces was extracted from structures of known protein–RNA complexes in the Protein Data Bank (PDB) [15] solved by X-ray crystallography. Proteins with >30% sequence identity or structures with resolution worse than 3.5 Å were removed using PISCES [16]. The resulting dataset, RB147 [36], contains 147 non-redundant polypeptide chains. RNA-binding residues were identified according to a distance-based cutoff definition: an RNA-binding residue is an amino acid containing at least one atom within 5 Å of any atom in the bound RNA. RB147 contains a total of 6157 RNA-binding residues and 26,167 non-binding residues. The RB147 dataset [36] is larger than the RB109 dataset used in our previous studies [17, 18].

DNA-protein interface dataset

A dataset of protein-DNA interfaces was extracted from structures of known protein-DNA complexes in the PDB [15]. Proteins with >30% sequence identity or structures with resolution worse than 3.0 Å and R factor > 0.3 were removed using PISCES [16]. The resulting dataset, DB208, contains 208 polypeptide chains, each at least 40 amino acids in length. DNA-binding residues were identified according to a definition based on reduction in solvent accessible surface area (ASA): an amino acid is a DNA-binding residue if its ASA computed in the protein-DNA complex using NACCESS [19] is less than its ASA in the unbound protein by at least 1 Å^2 [20]. DB208 contains a total of 5,721 interface residues and 39,815 non-interface residues. The DB208 dataset is larger than the DB171 dataset used in our previous studies [21].

2.2 Algorithms for predicting interfacial residues

We used sequenced-based Naïve Bayes classifiers [22, 23] for predicting protein-RNA interfaces [17, 18] and protein-DNA interfaces [21]. Briefly, the input to the classifier is a contiguous window of 2n+1 amino acid residues consisting of the target residue and n sequence neighbors to the left and right of the target residue, obtained from the protein sequence using the "sliding window" approach. The output of the classifier is a probability that the target residue is an interface residue given the identity of the 2n+1 amino acids in the input to the classifier. With Naïve Bayes classifiers, it is possible to tradeoff the rate of true positive predictions against the rate of false positive predictions, by using a classification threshold, θ, on the output probability of the classifier. The target residue is predicted to be an interface residue if its probability returned by the classifier is greater than θ, and a non-interface residue otherwise. The length of the window was set to 21 in the experiments described here.

We used the implementation of the Naive Bayes classifier available in WEKA, an open source machine learning package [23] for training classifiers used to predict interface residues in this study. The performance of the protein-RNA interface predictor trained on RB147 dataset (***RNABindR***, *http://bindr.gdcb.iastate.edu/RNABindR/*), and estimated using leave-one-out sequence-based cross-validation, is documented in [36]. The performance of protein-DNA interface predictor trained on the DB208 dataset (***DNABindR,*** *http://cild.iastate.edu/DNABindR*) and estimated using 10-fold sequence-based cross-validation, is comparable to that of the previously published protein-DNA interface predictor, which was trained on the DB171 dataset [21]. The RNA interface predictions on TEN domains were obtained by using Naïve Bayes classifiers trained on the RB147 dataset (high specificity setting of ***RNAbindR***). The DNA interface predictions were obtained by ***DNABindR*** (θ=0.168) trained on the DB208 dataset.

2.3 Structural modeling of telomerase TEN domains in human and yeast

The N-terminal domains from human telomerase (GENBANK NP_937986) and yeast telomerase (GENBANK NP_013422) sequences, were threaded onto the *T. thermophila* telomerase N-terminal domain (TEN) structure (PDB: 2b2a chain A) using FUGUE [24]. The output alignments were used for generating 3D coordinates for the N-terminal domains of human and yeast telomerase by MODELLER [25]. Among 15 generated models, the highest ranking model was chosen and refined using SCWRL [26] to reposition side-chains. Energy minimization was performed by 400 steps of steepest descent using the GROMOS96 force field [27] with a 9Å non-bonded cutoff in the Deep View/Swiss PDB-viewer [28]. One human TEN model was based on the *Tetrahymena* TEN structure in the PDB: 2b2aA, N-terminal domain of tTERT. For a second model, several templates were selected using PSI-BLAST [29] and the Swiss-Model HMM template library [30] to detect remote homologs of hTERT. The chosen templates were portions of the following PDB structures: 1imhC, Tonicity-responsive enhancer binding protein (TONEBP)-DNA complex; 1jfiB, Negative Cofactor 2-TATA box binding protein-DNA complex (NC2-TBP-DNA); 2dyrM, bovine heart cytochrome C oxidase; 1b1uA, bifunctional inhibitor of Trypsin and Alpha-amylase from Ragi seeds; 2b2aA, N-terminal domain of tTERT. The templates were aligned and models were generated using the procedure described above. All generated structures were evaluated using the ANOLEA server [34].

2.4 Experimental identification of RNA and DNA binding residues

Experimentally determined DNA and RNA binding sites in hTERT and tTERT were collected by mining relevant literature. Point mutations that affect RNA binding have not been reported, but Moriarty et al. showed that deletions at positions 30-39 and 110-119 in hTERT result in reduced RNA and DNA association, respectively [31, 32]. Conserved primer grip regions have been mapped in the TEN and RT domains of hTERT, between amino acids 137-141 and 930-934 [33]. Alanine substitutions in the C-terminal region of TEN at positions Q168, F178, and W187 have been shown to substantially decrease tTERT association with DNA [12].

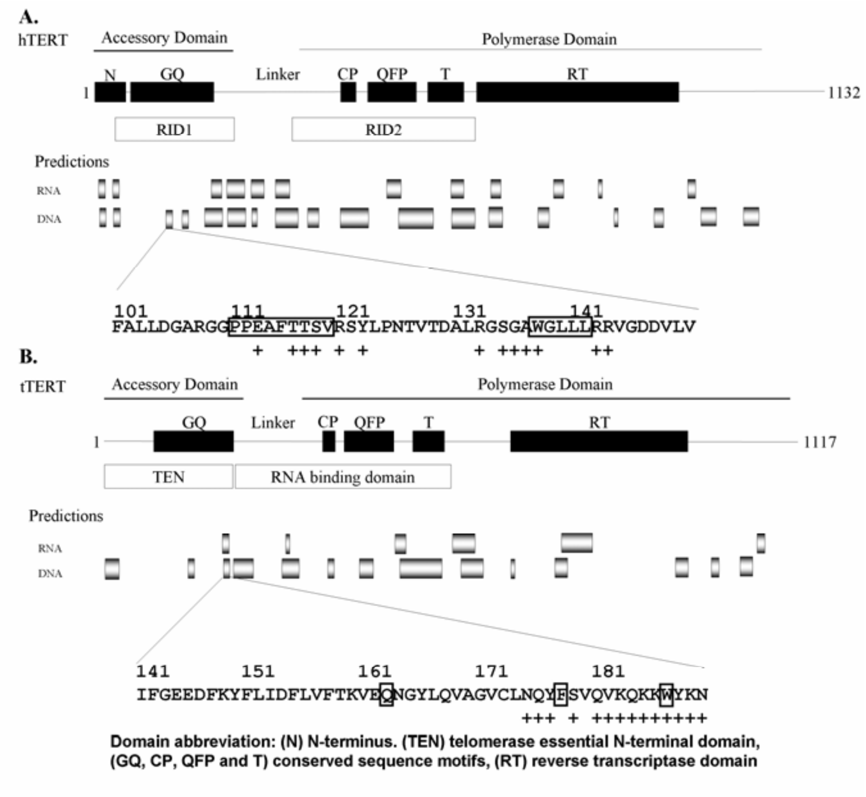
3. Results

3.1 Rationale

Computational and bioinformatic analyses can provide valuable insight into protein sequence-structure-function relationships, especially when the structure of a protein or complex is difficult to solve using experimental approaches. Surprisingly, despite the fascinating structural and regulatory complexity of telomerase, its pivotal role in cellular signal pathways, and its critical interactions with DNA, RNA and protein partners, very few studies have exploited bioinformatic or computational structural biology approaches to investigate the structure and function of telomerase. In this work, we use a combination of comparative structural modeling and sequence-based machine learning methods to test the hypothesis that the N-terminal domains of TERTs in diverse organisms share a similar overall architecture and conserved DNA and RNA binding surfaces.

3.2 Sequence-based prediction of RNA and DNA binding sites in human and Tetrahymena TERT

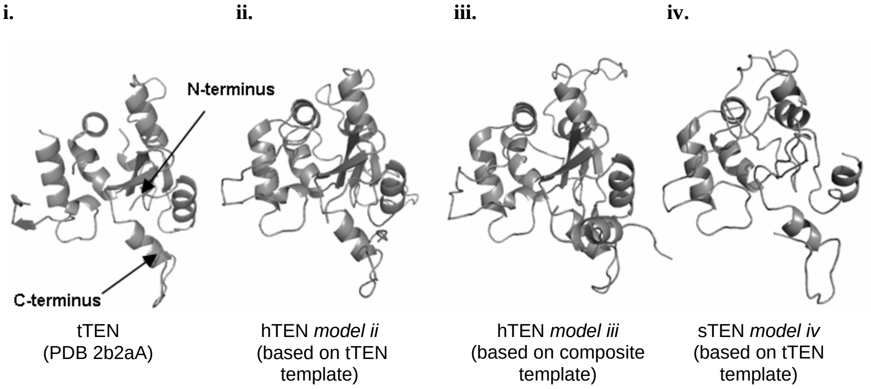
Conserved domains within the telomerase reverse transcriptase protein of human (hTERT) and *Tetrahymena* (tTERT) are illustrated in Figure 2. In previous work, we used a sequence-based machine learning approach to predict RNA binding residues in TERT sequences and showed that our predictions compared favorably with available experimental data [18]. Results of these previously published predictions are included in Figure 2 for comparison with DNA binding residues predicted in the current study (see Materials and Methods). The predicted DNA and RNA binding regions in hTERT and tTERT are indicated by boxes under the middle sections of Figures 2A and B, respectively. The lower portion of each figure shows specific examples, with boxed amino acids representing short deletions (in hTERT) or alanine-substitution mutations (in tTERT), that have been shown to compromise or abolish DNA binding. Note that for hTERT, the predictions either overlap or surround the amino acids implicated by deletion (Figure 2A). For tTERT, two of three experimentally-identified DNA binding residues lie within the predicted DNA binding region (Figure 2B).



Domain abbreviation: (N) N-terminus. (TEN) telomerase essential N-terminal domain, (GQ, CP, QFP and T) conserved sequence motifs, (RT) reverse transcriptase domain

Figure 2. Predicted interface residues and conserved domains for telomerase reverse transcriptase (TERT). Mapped functional domains and conserved motifs of TERT are shown above shaded boxes representing clusters of predicted RNA and DNA interface residues. Predicted interface residues are indicated by a + below the amino acid sequence. **A)** Human telomerase reverse transcriptase (hTERT). In the sequence shown, boxed amino acids 110-119 and 137-141, correspond to the template anchor site and a putative primer grip, implicated in forming the hTERT-DNA active complex [31, 32, 34]. **B)** *Tetrahymena* telomerase reverse transcriptase (tTERT). The amino acid sequence shown represents the C-terminal end of the TEN domain. Alanine mutations at positions Q168, F178 and W187 have been shown to significantly reduce hTERT-DNA association. Predicted interactions spanning amino acids 181-190 are located in a highly flexible, disordered region [12].

A.



B.

```
                                    •
T. thermophila     ----MQKINNINNNKQMLTRKEDLLTVLKQISALKYVSN--LYEFLLATEKIVQTSELDT
H. sapiens         ----MPRAPRCRAVRSLLRSHYREVLPLATFVRRLGPQG---WRLVQRGDPAAFRALVAQ
S. cerevisiae      --------------MKILFEFIQDKLDIDLQTNSTYKEN------LKCGHFNGLDEILTT
                                                                          •
T. thermophila     QFQEFLTTTII--ASEQNLVENYKQKYN-----QPNFSQLTIKQVID------DSIILLG
H. sapiens         CLVCVPWD-----ARPPPAAPSFRQVSC-----LKELVARVLQRLCE---RGAKNVLAFG
S. cerevisiae      CFALPNSR-------KIALPCLPGDLSH-----KAVIDHCIIYLLTG---ELYNNVLTFG
                                                               •       •
T. thermophila     NKQNY--VQQIGTTTIGFYVEYENINLSRQTLYSSNFRNLLNIFGEEDFKYFLIDFLVFT
H. sapiens         FALLDGARGGPPEAFTTSVRSYLPNTVTDALRGSGAWGLLLRRVGDDVLVHLLARCALFV
S. cerevisiae      YKIAR------NEDVNNSLFCHSAN-VNVTLLKGAAWKMFHSLVGTYAFVDLLINYTVIQ
                           •  •
T. thermophila     KVEQNGYLQVAGVCLNQYFSVQVKQKKWYKNN----
H. sapiens         LVAPSCAYQVCGPPLYQLGAATQARPPPHASGPRRR
S. cerevisiae      FNG-QFFTQIVGNRCNEPHLPPKWVQRSSSSSAT--
```

Figure 3. Comparison of TEN domain structures and sequences and in *Tetrahymena,* human and yeast, *S. cerevisiae.* **A)** Comparison of *Tetrahymena* TEN domain structure determined by X-ray crystallography with modeled structures of TEN domains from other species. **i)** *T. thermophila,* experimentally-determined structure, PDB ID: 2b2aA [12]; **ii)** human structural model, based on threading using the *T. thermophila* 2b2aA structure as template; **iii)** human structural model, based on threading using a composite of several different structures as template; **iv)** yeast, *S. cerevisiae*, structural model, based on threading using the *T. thermophila* 2b2aA structure as template. **B)** Multiple sequence alignment of telomerase TEN domains from *T. thermophila, H. sapiens,* and *S. cerevisiae* [12]. Amino acids conserved in all 3 species in the multiple sequence alignment are highlighted.

3.3 Structural modeling of N-terminal domain of TERT from human and yeast

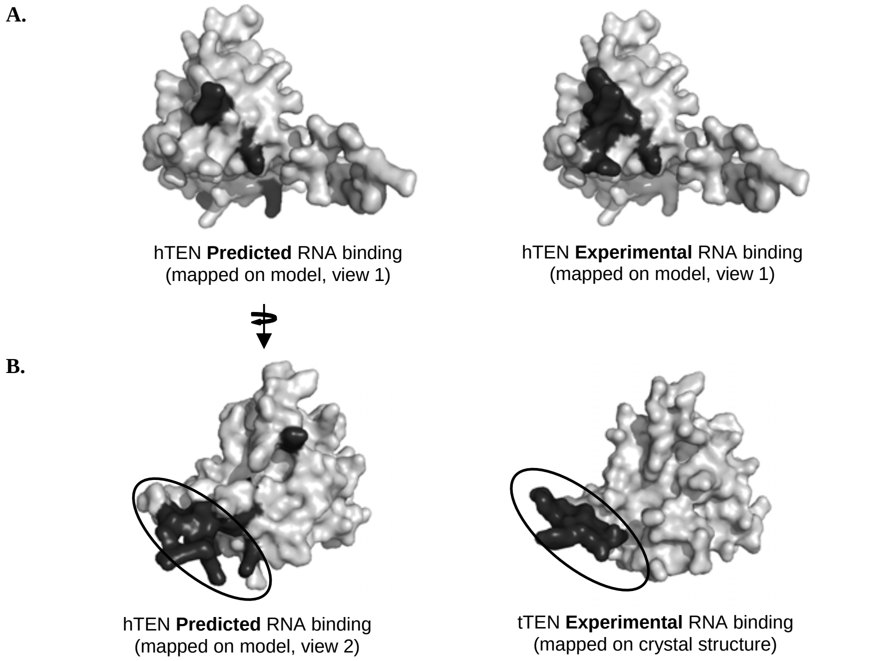
Our initial attempts to generate structural models of the human and yeast TEN domains by submitting their sequences to several web-based homology modeling servers were unsuccessful, due to failure of the servers to identify appropriate homology modeling templates (the pairwise sequence identity between TEN domains of hTERT and tTERT is < 20%). However, the results of multiple sequence alignment (Figure 3B) and predicted secondary structure similarities (data not shown), led us to try threading, using the FUGUE server (see Materials and Methods). The *Tetrahymena* TEN domain structure (PDB ID 2b2aA) was identified as the highest scoring structural template for both the human and yeast TEN domain sequences (hTERT: certain, with 99% confidence; sTERT: likely, with 95% confidence). Based on the alignments generated by FUGUE, we generated all-atom models and performed energy minimization to generate the final models illustrated in Figure 3A (see Materials and Methods for details). Two different models for the human TEN domain, *model ii*, based on the *Tetrahymena* TEN template, and *model iii*, based on a composite template from several different structures, were very similar to one another as well as to *model iv*, for the yeast TEN domain, despite their highly divergent amino acid sequences. Table 1 shows the root mean square deviation (RMSD) values calculated for comparison of the *Tetrahymena* TEN domain structure (determined by X-ray crystallography [12]) with the hTEN and sTEN modeled structures, using TOPOFIT [35] for structural alignment.

Aligned Structures	**RMSD (Å)**
tTEN *vs* hTEN	1.11
tTEN *vs* sTEN	1.41
sTEN *vs* hTEN	1.39

Table 1. RMSD computed from structural alignments of TEN domain structures: tTEN, *Tetrahymena,* PDB structure, 2b2aA (Fig.3A, *structure i)*; hTEN, human, modeled structure (Fig. 3A, *model ii*); sTEN, yeast, modeled structure (Fig. 3A, *model iv*). Alignments were performed using TOPOFIT [35]

3.4 Analysis of RNA and DNA binding surfaces in human and Tetrahymena TEN domains

To compare RNA and DNA binding surfaces in human and *Tetrahymena* TEN domains, we examined both our predicted nucleic acid binding sites and available experimental data in the context of the experimentally determined structure of *Tetrahymena* TEN domain [12] and modeled structure of the human TEN domain (*model ii*, Figure 3A). Examples of these analyses are illustrated in Figures 4 and 5. The predicted RNA binding residues in hTEN overlap with several RNA binding sites implicated by deletion experiments (Figure 4A, compare left and right models). Furthermore, additional putative RNA binding residues on the "back" side of the hTEN model (Figure 4B, left, in oval) co-localize with an experimentally defined RNA binding site mapped onto the tTEN crystal structure (Figure 4B, right, in oval).



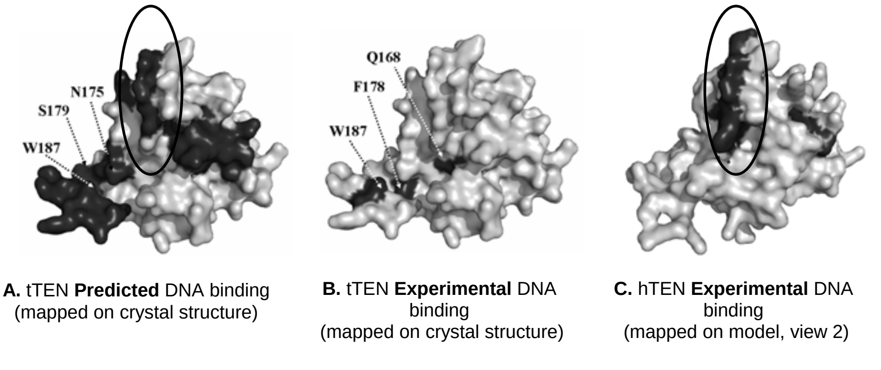
hTEN **Predicted** RNA binding
(mapped on model, view 1)

hTEN **Experimental** RNA binding
(mapped on model, view 1)

hTEN **Predicted** RNA binding
(mapped on model, view 2)

tTEN **Experimental** RNA binding
(mapped on crystal structure)

Figure 4. Comparison of predicted and experimentally determined RNA binding surfaces in TEN domains. A) Sequence-based RNA binding site predictions mapped onto the hTERT TEN domain *model ii* (left) overlap with experimentally determined RNA binding residues (right); Black residues are predicted (left) or actual (right) RNA binding residues. **B)** Another patch of predicted RNA binding residues in the hTEN model (left, in oval) co-localizes with an experimentally verified RNA binding region in tTEN (right). Figures 4 and 5 were generated using PyMol (http://pymol.sourceforge.net/).



A. tTEN **Predicted** DNA binding
(mapped on crystal structure)

B. tTEN **Experimental** DNA binding
(mapped on crystal structure)

C. hTEN **Experimental** DNA binding
(mapped on model, view 2)

Figure 5. Comparison of predicted and experimentally determined DNA binding surfaces in TEN domains. A) Residues predicted to interact with DNA (black), mapped onto tTEN, PDB 2b2aA. Predicted binding sites encompass residues shown in **B)** which illustrates the only 3 experimentally defined DNA binding residues in tTEN (see Fig. 2B). Note that additional predicted DNA binding residues in **A** (in oval) are consistent with **C)**, which shows experimentally validated DNA binding residues in the human protein mapped onto our modeled structure of hTEN.

Only three DNA binding residues in the TEN domain of tTERT have been experimentally identified: Q168, F178, and W187 (Figure 5B). Several additional putative DNA binding residues are predicted by our machine learning classifiers (Figure 5A). Some of these predicted residues in tTEN (in oval) co-localize with experimentally defined DNA binding residues in the human protein, when viewed in the context of our modeled structure of the hTEN domain (Figure 5C).

Taken together, these results support our hypothesis that TEN domains in diverse organisms have similar three dimensional structures and conserved nucleic acid binding surfaces. Further, they identify additional putative interface residues that could be targeted in experiment studies.

4. Summary and Discussion

Telomerase is one of several clinically important regulatory proteins for which it has been difficult to obtain high resolution structural information. The recent experimental determination of the structure of the N-terminal domain of tTERT, the telomerase reverse transcriptase component from *Tetrahymena*, suggests that at least partial structural information for human telomerase may soon become available. It seems unlikely, however, that experimental elucidation of the structure of the multisubunit RNP complex corresponding to the catalytically active form of telomerase will occur in the near future. Thus, the integrative strategy proposed here, in which structural information gleaned from comparative modeling is combined with machine learning predictions of functional residues, can be expected to provide valuable insights into the sequence and structural correlates of function for telomerase and other "recalcitrant" proteins. We are currently pursuing several avenues for improving the reliability of machine learning predictions, including the use of different sequence representations and additional sources of input information (e.g., structure and phylogenetic information, when available) and more sophisticated machine learning algorithms. We are also pursuing additional approaches for protein structure prediction, including ab initio and fold recognition methods capable of incorporating predicted protein-protein contacts as constraints. Given the large number of proteins with which telomerase interacts and the essential roles of telomerase in cellular signaling, aging, cancer, and other human diseases, this should continue to be rich and challenging area of research.

5. Acknowledgements

This research was supported in part by NIH GM 066387, NIH-NSF BSSI 0608769, NSF IGERT 0504304 and by the ISU Center for Integrated Animal Genomics. We thank Fadi Towfic for critical comments on the manuscript and members of our groups for helpful discussions.

References

1. E. H. Blackburn, *FEBS Letters* **579**, 859 (2005).
2. K. Collins, *Nat. Rev. Mol. Cell. Biol.* **7**, 484 (2006).
3. C. Autexier and N. F. Lue, *Annu. Rev. Biochem.* **75**, 493 (2006).
4. C. W. Greider and E. H. Blackburn, *Cell* **43**, 405 (1985).
5. E. H. Blackburn, *Mol. Cancer. Res.* **3**, 477 (2005).
6. J. W. Shay and W. E. Wright, *J. Pathol.* **211**, 114 (2007).
7. M. A. Blasco, *Nat. Rev. Genet.* **8**, 299 (2007).
8. T. de Lange, Genes. Dev. **19**, 2100 (2005).
9. S. B. Cohen, M. E. Graham, G. O. Lovrecz, et al., *Science* **315**, 1850 (2007).
10. N. Hug and J. Lingner, *Chromosoma* **115**, 413 (2006).
11. A. Goldkorn and E. H. Blackburn, *Cancer Res.* **66**, 5763 (2006).
12. S. A. Jacobs, E. R. Podell, T. R. Cech, *Nat.Struct.Mol.Biol.* **13**, 218 (2006).
13. K. L. Friedman and T. R. Cech, *Genes Dev.* **13**, 2863 (1999).
14. J. Xia, Y. Peng, I. S. Mian, et al., *Mol. Cell. Biol.* **20**, 5196 (2000).
15. H.M. Berman, J. Westbrook, Z. Feng, et al., *Nucleic Acids.Res.* **28**, 235 (2000).
16. G. Wang and R. L. Dunbrack, Jr., *Bioinformatics* **19**, 1589 (2003).
17. M. Terribilini, J. H. Lee, C. Yan, et al., *Pac. Symp. Biocomput.*, 415 (2006).
18. M. Terribilini, J. H. Lee, C. Yan, et al., *RNA* **12**, 1450 (2006).
19. S. J. Hubbard, S. F. Campbell, J.M. Thornton, *J. Mol. Biol.* **220**, 507 (1991).
20. S. Jones and J. M. Thornton, *Proc. Natl. Acad. Sci.* **93**, 13 (1996).
21. C. Yan, M. Terribilini, F. Wu, et al., *BMC Bioinformatics* **7**, 262 (2006).
22. T. Mitchell, *Machine Learning* (McGraw-Hill, 1997).
23. I.H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques* (Morgan Kaufmann, 2005).
24. J. Shi, T. L. Blundell, and K. Mizuguchi, *J. Mol. Biol.* **310**, 243 (2001).
25. R. Sanchez and A. Sali, *Proteins* **Suppl 1**, 50 (1997).
26. A. A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack, Jr., *Protein Sci.* **12**, 2001 (2003).
27. W. R. P. Scott, P. H. Hunenberger, I. G. Tironi, et al., *J. Phys. Chem.* **A 103**, 3596 (1999).
28. N. Guex and M. C. Peitsch, *Electrophoresis* **18**, 2714 (1997).
29. S. F. Altschul, T. L. Madden, A. A. Schaffer, et al., *Nucleic Acids. Res.* **25**, 3389 (1997).
30. J. Kopp and T. Schwede, *Nucleic Acids. Res.* 32, D230 (2004).
31. T. J. Moriarty, S. Huard, S. Dupuis, et al*., Mol. Cell. Biol.* **22**, 1253 (2002).
32. T.J. Moriarty, R.J. Ward, M.A. Taboski, et al., *Mol.Biol.Cell.* **16**, 3152 (2005).
33. H. D. Wyatt, D. A. Lobb, and T. L. Beattie, *Mol. Cell. Biol.* **27**, 3226 (2007).
34. F. Melo and E. Feytmans, *J. Mol. Biol.* **277**, 1141 (1998).
35. V. A. Ilyin, A. Abyzov, and C. M. Leslin, *Protein Sci.* **13**, 1865 (2004).
36. M. Terribilini, J. D. Sander, J. H. Lee, et al., *NAR* **35**, W578-W584 (2007).