Early Vertebrate Evolution of the TATA-Binding Protein, TBP

Alla A. Bondareva and Edward E. Schmidt
Veterinary Molecular Biology, Marsh Labs, Montana State University

TBP functions in transcription initiation in all eukaryotes and in Archaeabacteria. Although the 181-amino acid (aa) carboxyl (C-) terminal core of the protein is highly conserved, TBP proteins from different phyla exhibit diverse sequences in their amino (N-) terminal region. In mice, the TBP N-terminus plays a role in protecting the placenta from maternal rejection; however the presence of similar TBP N-termini in nontherian tetrapods suggests that this domain also has more primitive functions. To gain insights into the pretherian functions of the N-terminus, we investigated its phylogenetic distribution. TBP cDNAs were isolated from representative nontetrapod jawed vertebrates (zebrafish and shark), from more primitive jawless vertebrates (lamprey and hagfish), and from a prevertebrate cephalochordate (amphioxus). Results showed that the tetrapod N-terminus likely arose coincident with the earliest vertebrates. The primary structures of vertebrate N-termini indicates that, historically, this domain has undergone events involving intragenic duplication and modification of short oligopeptide-encoding DNA sequences, which might have provided a mechanism of de novo evolution of this polypeptide.

Introduction

The C-terminal core of the TATA-binding protein (TBP) is one of the most evolutionarily conserved protein domains known; however the TBP N-terminal region varies greatly between taxa (Hernandez 1993). Published sequences indicate that all tetrapods (amphibians, reptiles, birds, and mammals) have similar TBP N-termini (Hernandez 1993; Sumita et al. 1993; Nakashima et al. 1995; Shimada et al. 1999). Conversely, TBP sequences from nonvertebrate metazoans, including both protostomes (e.g., arthropods) and lower deuterostomes (e.g., echinoderms), lack nonrepetitive sequences resembling the tetrapod TBP N-terminus (NCBI/GenBank) (Lichtsteiner and Tjian 1993; Muhich et al. 1990). The tetrapod TBP N-terminus can be divided into four subdomains (fig. 1a). Two of these, entitled N_N and N_C, which do not resemble each other or other known proteins, flank a central glutamine-rich repeat region, called Q. Near the junction with the conserved C-terminus is an imperfect repeat region of sequence (PXT)n, where X is generally M, A, or I. This repeat exists in most or all metazoan TBPs, and we suspect it serves as a connector that integrates phyla-specific properties of the various metazoan N termini with the universal functions of the C-terminus.

We hypothesized that the N-terminus functions as a “signaling port” by which the basal transcription machinery receives specific regulatory signals and that variation within this region represents commitment to gene-regulation processes that are unique to each phylum. We previously showed that mice lacking most of this domain exhibit no defects in basal cellular/metabolic functions nor in any other functions shared between mammals and nonvertebrate eukaryotes (Hobbs et al. 2002; Schmidt et al. 2003), as one might predict based on the absence of related sequences from lower eukaryotes. Rather, homozygous mutants exhibit a lethal defect in the function of the midgestational placenta, which can be rescued by providing mutant fetuses with a wild-type tetrapolid placenta. Genetic rescue studies suggest mutant failure results from perturbation of a β2-microglobulin–dependent process that the placenta uses to evade maternal immune rejection (Hobbs et al. 2002).

The physiological site of the defect in these mice, the chorionic placenta, is found only in therian mammals. However, the presence of a similar TBP N-terminus among all tetrapods suggests this domain confers a fitness advantage on nonplacental as well as placental tetrapods. To gain insights into the preplacental functions of this domain, we isolated TBP cDNAs from strategically placed representatives of nontetrapod vertebrate lineages (teleost, shark, lamprey, and hagfish) and from an advanced nonvertebrate deuterostome (amphioxus). Within the resolution afforded by analyses of extant species, this domain was precisely restricted to vertebrates. Comparisons between species revealed that multiple events involving intragenic duplication of oligopeptide-encoding DNA sequences have occurred and persisted in this domain.

Materials and Methods

Isolation of cDNA and Genomic tbp Clones

Sources cited in the acknowledgments provided cDNA libraries from amphioxus (Branchiostoma floridae) (Langeland et al. 1998), Atlantic sea lamprey (Petromyzon marinus) (Tomsa and Langeland 1999), nurse shark (Ginglymostoma cirratum) (Kandil et al. 1996), Pacific hagfish (Eptatretus stoutii) (Nagata et al. 2002), and zebrafish (Danio rerio). A cDNA λ library was produced in a ZAP II vector using poly-A+ mRNA isolated from spawning Atlantic sea lamprey testes. The library contained 2.25 × 10^6 primary clones; 84% contained inserts of approximately 1.5 kb average length.

For screening zebrafish and shark libraries, a cDNA probe containing most of the C-terminus and 222 bp of 3’ untranslated region (UTR) of mouse (m) TBP was used. For screening lamprey and amphioxus libraries (8 × 10^5 to 10 × 10^5 phage screened per library), we used a mixed-species probe containing the same mTBP cDNA fragment and (1) an 840-bp Hinc II/Sca I fragment of shark TBP cDNA (144 bp of N-terminus, entire C-terminus, and 149 bp of 3’ UTR); and (2) a 697-bp fragment of zebrafish TBP cDNA (start codon to a position midway through the
C-terminus). TBP-containing inserts were recombined into SK-plasmid vectors and inserts were sequenced in both directions. Five amphioxus, two lamprey, two shark, and five zebrafish TBP cDNA clones (designated aTBP, lTBP, sTBP, and zfTBP, respectively) were sequenced.

We isolated 23 TBP clones from 1.8 \times 10^6 plaques of the hagfish library (designated hfTBP); however none were full length. The two longest cDNAs encoded the entire C-terminus, but only 91 aa of the N-terminal region. The sequence revealed PstI, HindIII, and EcoRI sites in the 5'9 end of the C-terminal protein-coding DNA. Genomic Southern blots suggested that the hagfish gene also contained PstI, EcoRI, and HindIII sites at approximately 2.2 kb, approximately 3.5 kb, and approximately 6.7 kb upstream of the TBP N-terminal/C-terminal junction, respectively. Using Hind III ligation-mediated PCR (hfTBP-specific primer: 5'-TAT GGA TCC TGA CGT TGT GCT TCC ACT G-3'), we isolated and cloned a 1,359-bp genomic fragment, including more of the TBP N-terminal region. This fragment contained two exons, which, together, encoded a protein domain homologous to the entire exon 3-encoded N-terminal region of mTBP. It did not, however, include the region homologous to mTBP exon 2, (start codon and subsequent 17 aa) (Sumita et al. 1993; Ohbayashi et al. 1996). A primer to sequences from the 5' protein-coding region of this clone (5'-TAT GGA TCC AGC ATC TAG TTG GTT CTG CC-3') was used in combination with the vector-encoded T3 primer to screen 24 portions (1 \times 10^6 phage/portion) of the hagfish library by PCR. Two portions contained full-length hfTBP cDNAs, including 200 bp of 5' UTR.

Sequence Alignments, \(d_N\) and \(d_S\) Estimations, and Relatedness Predictions

Initial pair-wise alignments were performed using Blast (NCBI/GenBank). Progressive multiple sequence alignments were performed using ClustalW software (Thompson, Higgins, and Gibson 1994) and best-fit/gap-placement was confirmed manually (fig. 1) (Thompson, Higgins, and Gibson 1994). \(d_S/d_S\) estimations were performed with the MEGA version 2.1 software package (Kumar et al. 2001) using five methods (Li, Wu, and Luo 1985; Nei and Gojobori 1986; Li 1993; Pamilo and
Table 1

<table>
<thead>
<tr>
<th>Polypeptide</th>
<th>Codons</th>
<th>Accession Numbers</th>
<th>Method 1</th>
<th>Method 2</th>
<th>Method 3</th>
<th>Method 4</th>
<th>Method 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>TBP N-terminus</td>
<td>114</td>
<td>D01034, D83127, D31776, X66033</td>
<td>0.03745</td>
<td>0.05385</td>
<td>0.04535</td>
<td>0.04742</td>
<td>0.05039</td>
</tr>
<tr>
<td>TBP C-terminus</td>
<td>181</td>
<td>D01034, D83127, D31776, X66033</td>
<td>0.00588</td>
<td>0.00741</td>
<td>0.00696</td>
<td>0.00791</td>
<td>0.00936</td>
</tr>
<tr>
<td>Histone H2B</td>
<td>125</td>
<td>NM_030082, D70896, X03018</td>
<td>0.07043</td>
<td>0.08456</td>
<td>0.06683</td>
<td>0.06594</td>
<td>0.07875</td>
</tr>
<tr>
<td>Translation eEF-2</td>
<td>857</td>
<td>BC007152, U66663, BC044327</td>
<td>0.02910</td>
<td>0.02086</td>
<td>0.02952</td>
<td>0.02274</td>
<td>0.02692</td>
</tr>
<tr>
<td>TBP C-terminus</td>
<td>58</td>
<td>D01034, D83127, D31776, X66033</td>
<td>0.01839</td>
<td>0.02557</td>
<td>0.02183</td>
<td>0.02085</td>
<td>0.02392</td>
</tr>
<tr>
<td>Translation H2B</td>
<td>959</td>
<td>NM_007392, M10607, AF416707, X12525</td>
<td>0.01839</td>
<td>0.02392</td>
<td>0.02085</td>
<td>0.02183</td>
<td>0.02557</td>
</tr>
</tbody>
</table>

&d for nonsynonymous substitutions from published tetrapod TBP C-termini gave values between 0.006 and 0.009, which is roughly threefold lower than those calculated for \( \alpha \)-actin, fivefold lower than those for translation elongation factor 2 (eEF-2), and 10-fold lower than those for histone H2B (table 1). In the absence of selection, a \( ds/dN \) of approximately 1.0 would be expected (reviewed in Nei and Kumar 2000, Yang 2001, and Yang 2002). The low \( ds/dN \) values for the tetrapod TBP C-termini suggest that this domain has been subjected to extraordinarily strong purifying selection (Nei and Kumar 2000; Yang 2001; Yang 2002). However, greater variation is seen within the N-terminus (Tamura et al. 1991). Thus, six substitutions per 100 aa per species were counted between the same four tetrapods (data not shown), and \( ds/dN \) values of 0.04 to 0.06 were obtained (table 1), which were comparable to those of H2B.

Although these results suggested that the TBP N-terminus was as conserved between tetrapods as was H2B, published data indicated that nonvertebrate species, including nematodes, insects, and echinoderms, lacked sequences resembling the tetrapod N-terminus (NCBI/GenBank) (Muhich et al. 1990; Tamura et al. 1991; Lichtsteiner and Tjian 1993). Therefore, we initiated a study aimed at determining which species between echinoderms and tetrapods contained sequences related to the tetrapod N-terminus.

Phylogenetic Distribution of the N-terminus

TBP cDNAs were isolated from a cephalochordate (amphioxus), from two primitive jawless vertebrates (hagfish and lamprey), from a jawed cartilaginous fish (nurse shark), and from a teleost (zebrafish). Three TBP mRNA isoforms, encoding two protein isoforms, were isolated from amphioxus. The two proteins differed only by a QG insertion/deletion within an XQ heterodipeptide repeat region (fig. 1b). In addition, three different 3' UTR isoforms were found, which differed by the presence or absence of oligonucleotide insertions (fig. 1c). Thus, as compared with the shortest isoform (number 1), isoform number 2 had four insertions of lengths 7, 8, 18, and 23 bp, respectively. Isoform number 3 lacked these insertions, but
had two other insertions of lengths 4 and 7 bp. The 8-bp insertion in the former isoform, and the 7-bp insertion in the latter, are direct repeats of adjacent sequences; other oligonucleotide insertions resemble nearby sequences, suggesting that they might be diverged repeats.

The amphioxus N-terminus, like those of all metazoans, had a PXT repeat; however, outside of this region, it lacked sequences suggesting homology with either the tetrapod N terminus (fig. 1b) or with TBP domains in any more primitive metazoans (NCBI/GenBank [data not shown]). Conversely, zebradish, shark, lamprey, and hagfish TBP N-termini resembled those of tetrapods (fig. 1b). Both within the C-terminal domain and within the N terminal, hagfish and lamprey TBP proteins had shared amino acids that differed from those in higher vertebrates (fig. 1b). Phylogenetic relationship predictions based on either the N-terminus or the C-terminus grouped lamprey and hagfish within a clade and separated vertebrates from either the N-terminus or the C-terminus. Relatedness is depicted as trees, with the p-distance scale indicated below each tree. The N-terminal sequences for B. floridae, D. melanogaster, and C. elegans cannot be reliably aligned with each other or with vertebrate TBP sequences, and thus these branches do not connect to the vertebrate tree or with each other (a) (see Materials and Methods).

We find no evidence that the N-termini from these four groups are homologous and an interconnected rooted tree is depicted (b).

Phylogenetic Distribution of Introns Within Sequences Encoding the N-Terminus

The positions of introns within genes are highly conserved during evolution (Long, de Souza, and Gilbert 1995; de Souza, Long, and Gilbert 1996; Gilbert, de Souza, and Long 1997). Although the TBP N-terminus in tetrapods is encoded almost entirely by a single large exon (Nakashima et al. 1995; Ohbayashi et al. 1996; Shimada et al. 1999), our analyses of *tbp* genomic sequences revealed an unexpected “extra” intron in this region of the hagfish *tbp* gene (see Materials and Methods). Analysis of genomic DNAs from the other vertebrates in our study indicated that lamprey also had an intron at this position; however, zebradish and shark did not (fig. 3). This intron forms a phylogenetic marker that distinguishes hagfishes and lampreys, as a group, from all other vertebrates.
The C-terminal 181-aa core of TBP is found from Archaea to man and may be the most highly conserved protein motif in the living world. The fusion of this evolutionarily static protein motif to the phyla-specific N-terminus provided a convenient baseline control for analyses of evolutionary changes within the N-terminus. Unexpectedly, the sequences of hfTBP and lTBP revealed the apparently repetitive nature of the ancestral N-terminal domain (fig. 1b) and the presence of the cyclostome-specific intron (fig. 3). The findings presented here lend insights into our understanding of early vertebrate phylogeny, the mechanisms by which a novel polypeptide motif like the TBP N-terminus might evolve, and the physiological functions of this domain.

Vertebrate Phylogeny and the TBP N-Terminus

Although the phylogenetic position of hagfishes has been controversial (Janvier 1999; Neidert et al. 2001; Delarbre et al. 2002), a recent report based on mitochondrial genomes concluded that hagfishes and lampreys share a clade, "cyclostomata," within vertebrata (fig. 4) (Delarbre et al. 2002). Relatedness prediction algorithms for both the N-terminal and the C-terminal regions of TBP group the hagfish and lamprey polypeptides within a separate vertebrate clade (fig. 2). Moreover, hagfishes and lampreys, but not jawed vertebrates, have an intron at an identical position within the region encoding the N-terminus (fig. 3). Because the entire protein-coding regions of these genes are homologous to those of higher vertebrates, the designations for conserved (i.e., homologous) exons and introns should be retained. Therefore, we have designated this new intron "C" for "cyclostomata," the clade to which this intron is restricted. The exons on either side of intron C are designated "3-left" and "3-right" to indicate that they are homologous to the left and right sides of exon 3 from higher vertebrates, respectively. We cannot determine whether intron C arose immediately after divergence of the lineage leading to cyclostomes or was more primitive but was lost from the lineage leading to jawed vertebrates (fig. 4). Nevertheless, in combination, our TBP sequence data and the presence of the cyclostome-specific intron C support the proposed monophyletic origin of cyclostomes (Delarbre et al. 2002).

We show that all vertebrates share homologous TBP N-termini, whereas the domain from amphioxus is dissimilar (fig. 1b). Thus, the vertebrate TBP N-terminus either arose coincident with the first vertebrates, or it immediately predated their appearance. Although studies on mice lacking this domain have not yet provided evidence that N-terminus–dependent processes played a causal role in the genesis of vertebrates (Hobbs et al. 2002; Schmidt et al. 2003), the precise vertebrate distribution and strong amino acid conservation of this domain suggest it might perform critical vertebrate-specific functions (see below).

Evolution of the Vertebrate TBP N-Terminus

The results presented here suggest the TBP N-terminus has undergone multiple events involving duplication of oligopeptide-encoding domains (fig. 1b). Expansion of heterodipeptide or larger repeats would
likely require duplications of 6-bp, 9-bp, or larger minisatellite sequences. The 3’ UTR sequences of individual TBP cDNAs from a polyclonal amphioxus population suggested that insertion/deletion of small minisatellite sequences occurs frequently in the 

\[ \text{tbp} \]

gene (fig. 1c). We also documented duplications in sequences encoding homooligopeptides (Q region), heterodipeptides (GQ polymorphism in amphioxus), heterotripeptides (PXT repeats), and larger heterooligopeptides (\(N_N\) and \(N_C\) repeats). A previous study noted the presence of several other 6-aa to 9-aa oligopeptide repeats within the TBP N-terminus of \(C.\ \text{elegans}\), \(Drosophila\), and humans, and suggested that these repeats might function in transcription (Lichtsteiner and Tjian 1993). Here we propose a model by which duplication and divergence of oligopeptide units might have allowed de novo evolution of this vertebrate-specific polypeptide domain.

It is generally accepted that two major routes by which novel genes appear is by duplication and divergence of existing genes or by duplication, “shuffling,” and divergence of existing exons (reviewed in Gilbert, de Souza, and Long 1997). Much less is known about the evolution of novel proteins or protein domains by other mechanisms. Our data suggest that, at or about the time of the appearance of the first vertebrates, the TBP N-terminus might have evolved de novo by duplications and divergence of minisatellite-encoded oligopeptide domains. Minisatellites are known to be major contributors to repetitive genomic regions (Heringa 1998). They have also been implicated in creating the repetitive polymorphisms within the mucin genes (Fowler, Vinall, and Swallow 2001). We theorize that duplication and divergence of minisatellite repeats can also contribute to the genesis of novel polypeptide domains.

Regions of arbitrary length might be duplicated during replication, as has been proposed as a general mechanism for minisatellite expansion (Levinson and Gutman 1987). Indeed, only one of the six minisatellite insertions that we found in the amphioxus 3’ UTRs (noncoding) was a multiple of 3 nt in length (fig. 1c). In protein-coding regions, modifications that altered the reading frame would disrupt the downstream TBP C-terminus and therefore would not persist. As a result, neither the nucleic acid sequence nor the amino acid sequence would need to have a deterministic effect on the reiterated mechanism, either in terms of reading-frame maintenance or in terms of the positions or lengths of the duplications. Rather, disruptive duplications or deletions would reduce fitness and would be lost from the lineage by natural selection. Modeling reveals that intragenic minisatellite duplications tend to tandemly reiterate amino acid motifs rather than introduce novel amino acids (fig. 5a).

Once dimerized, subsequent duplications of the same size will add another repeat of the exact same unit even if they initiate out-of-step with the existing repeats (fig. 5b). The propensity of this mechanism to duplicate short oligopeptide-encoding domains suggests it could play an important role in creating many reiterated secondary structures, such as the B-ZIP domain (7-aa diverged-repeat helix) (Landschultz, Johnson, and McKnight 1988), the ankyrin repeat (33-aa diverged fold structure) (Bork 1993; Mosavi, Minor, ...
and Peng 2002), the collagen repeat (GXP repeat), the C-terminal domain (CTD) repeat on the large subunit of RNA pol II (7-aa repeat) (Ahearn et al. 1987), and others.

Minisatellite duplications might also help explain why in TBP, unlike in huntingtin, both CAA and CAG codons are present in the Q region of most vertebrate species. Previously, length polymorphisms within the TBP Q region, like that within the Huntingtin Q region, were proposed to result from trinucleotide microsatellite expansions (Rubinsztein et al. 1996; Koide et al. 1999). If, instead, sequences within this region expanded as multicond codon minisatellites, as would be the case for the larger repeats within the N-terminus, then both Q codons might cotemplate individual duplication events. This could help account for the persistence of both codons in the Q region of most species.

Functions of the TBP N-Terminus

Because all vertebrates, but no nonvertebrates, possess a TBP N-terminus that is globally homologous to that of tetrapsids, this domain likely first appeared in an ancestor shared by all extant vertebrates but not by amphioxus (fig. 2). This suggests the vertebrate TBP N-terminus likely participates in functions that arose coincident with the earliest vertebrates and that it has imparted a fitness advantage on all subsequent vertebrate lineages.

Anatomically, the transition to vertebrates is correlated with the appearance of characteristics such as vertebrae, neural crest tissue, placodes, and others (Neidert et al. 2001; Delarbre et al. 2002). This transition was also associated with gene duplications leading to novel members of some developmental transcription factor families, including the Otx and Dlx families (Tomsa and Langeland 1999; Neidert et al. 2001), which participate in neural crest formation (Dolle, Price, and Duboule 1992; Vignali et al. 2000). However, the pathways in which these factors function (Simeone, Puelles, and Acampora 2002), as well as all of the known regulators and pathways that define neural crest (Baker and Bronner-Fraser 1997; Langeland et al. 1998; Neidert et al. 2001; Garcia-Castro, Marcelle, and Bronner-Fraser 2002; Knecht and Bronner-Fraser 2002), are far more ancient than is neural crest, per se. In the absence of a truly “vertebrate-specific” pathway regulating neural crest formation, it has been theorized that vertebrate-specific regulation of preexisting pathways allowed formation of this tissue uniquely in the vertebrate lineage (Baker and Bronner-Fraser 1997). In this light, it is interesting to consider that our work suggests the promotors of all genes in all vertebrates, but not in nonvertebrate phyla, are bound by TBP molecules that contain a unique N-terminal domain. In other words, all vertebrate transcription initiation complexes are “vertebrate-specific, not because of the DNA sequence of their promotors, but rather, because of this novel polypeptide domain within the basal transcription machinery.

The physiological functions for which natural selection has conserved this domain may differ between vertebrate phyla. Thus, although disruption of the TBP N-terminus in mice leads to a defect in therian pregnancy (Hobbs et al. 2002), no other vertebrates share this process, and, by inference, natural selection must have conserved this domain for other functions in pretherian vertebrates. These observations are consistent with model for evolution of this protein domain that we proposed based on functional studies in the mutant mice in which existing motifs can acquire novel activities that might later become “backed-up” to make the system robust (Schmidt et al. 2003). Since mice lacking the N-terminus exhibit only a defect in an evolutionary recent function, one might hypothesize that, if the N-terminus retains any of its more ancestral functions in mice, these can be accomplished by redundant mechanisms.

Acknowledgments

We thank K. Daughenbaugh, J. Prigge, D. Schwartzenberger, and M. Pratt in our group for technical assistance; A. Rzhetsky at Columbia University for consultation and advice; W. Swink at the Hammond Bay Biological Station for providing lampreys; J. Seaborne and D. Headlee at the Oregon Department of Fish and Wildlife for providing hagfish; and J. Langeland at Kalamazoo College, M. Flajnik at the University of Maryland, and D. Grunwald at the University of Utah for libraries and reagents. This work was supported by a Basil O’Connor Award from the March of Dimes Foundation, a Montana Agricultural Experimental Station appointment, and a grant from the National Science Foundation to EES. MAES/MSU/COA manuscript #2002-01.

Literature Cited


William Taylor, Associate Editor

Accepted June 21 2003