Evolutionary Analysis of the Mammalian Tuftelin Sequence Reveals Features of Functional Importance

S. Delgado, D. Deutsch, J. Y. Sire

HAL Id: hal-01516886

https://hal.sorbonne-universite.fr/hal-01516886

Submitted on 2 May 2017

Evolutionary analysis of the mammalian tuftelin sequence reveals features of functional importance

S. Delgado[1*], D. Deutsch[2] and J.Y. Sire[1]

[1]Evolution et Développement du Squelette, UMR7138- Evolution Paris-Seine, Institut de Biologie (IBPS), Université Pierre et Marie Curie, Paris, France.

[2] Dental Research Laboratory, Faculty of Dental Medicine, Institute of Dental Sciences, The Hebrew University of Jerusalem-Hadassah, Jerusalem, Israel.

* corresponding author: sidney.delgado@upmc.fr

Running title: Evolutionary analysis of TUFT1

Keywords

Tuftelin, TUFT1, MYZAP, Evolution, Mineralization, Mammals

Abstract

Tuftelin (TUFT1) is an acidic, phosphorylated glycoprotein, initially discovered in developing enamel matrix. TUFT1 is expressed in many mineralized and non-mineralized tissues. We performed an evolutionary analysis of 82 mammalian TUFT1 sequences to identify residues and motifs that were conserved during 220 million years (Ma) of evolution. We showed that 168 residues (out of the 390 residues composing the human TUFT1 sequence) are under purifying selection. Our analyses identified several, new, putatively functional domains, and confirmed previously described functional domains, such as the TIP39 interaction domain, which correlates with nuclear localization of the TUFT1 protein, that was demonstrated in several tissues. We also identified several sites under positive selection, which could indicate evolutionary changes possibly related to the functional diversification of TUFT1 during evolution in some lineages. We discovered that TUFT1 and MYZAP (myocardial zonula adherens protein) share a common ancestor that was duplicated circa 500 million years ago. Taken together, these findings expand our knowledge of TUFT1 evolution and provide new information that will be useful for further investigation of TUFT1 functions.

INTRODUCTION

When initially discovered in developing and mature bovine enamel, the acidic, phosphorylated glycoprotein tuftelin (TUFT1) was thought to play a major role in the structural organization and mineralization of enamel (Deutsch 1989; Deutsch et al. 1991). Subsequent studies focused on its putative function during enamel formation in various mammalian species (Deutsch et al. 1991, 1998; Bashir et al. 1998; McDougall et al. 1998). The protein was originally identified in ameloblasts and in the extracellular matrix at the dentine-enamel junction (DEJ), and its pattern and timing of expression during amelogenesis suggested that TUFT1 could be involved in the initial stages of enamel mineralization (Deutsch et al. 1991, 1998). The lack of a signal peptide at its N-terminus was, however, intriguing (Paine et al. 2000). In fact, TUFT1 is principally located in the Tomes' processes of secretory ameloblasts, that are shed during amelogenesis, and this may explain the presence of TUFT1 in the enamel matrix. In ameloblasts, TUFT1 was shown to form a complex with the Tuftelin-Interacting-Protein 39 (TIP39) (Paine et al. 2000). Several alternatively spliced tuftelin transcripts have been detected in humans and mice (Mao et al. 2001; Deutsch et al. 2002). Most transcripts lack a single exon but some may be missing several exons. The function of the resulting isoforms is still unknown. An analysis of SNPs in the *TUFT1* gene in a Turkish population suggested that some mutations were associated with a high frequency of caries, supporting the role of TUFT1 in enamel mineralization (Patir et al., 2008). Moreover, further studies indicated that *TUFT1* may be involved in individual predisposition to tooth hypomineralization (Jeremias et al. 2013).

Ten years after its discovery, it was demonstrated that TUFT1 expression was ubiquitous; TUFT1 is expressed in many mineralized and non-mineralized tissues, and in cancer cells (Mao et al. 2001; Deutsch et al. 2002; Leiser et al. 2007). These findings suggest that TUFT1 could have various fundamental roles that are not restricted to tooth

amelogenesis. Indeed, more recently, it was demonstrated that the expression level of TUFT1 mRNA was significantly higher in tissues in which oxygen levels hover closely to hypoxia under normal conditions (Leiser et al. 2007). In mouse brain, in a mouse mesenchymal C3H10T1/2 stem cell model, and in a neuronal PC12 cell model, TUFT1 expression was found to be induced by hypoxia via HIF1a (Leiser et al. 2010; Deutsch et al. 2011). During NGF-mediated PC12 differentiation, TUFT1 expression was significantly induced in correlation with neurite outgrowth, and partially blocked by K252a, a selective antagonist of the NGF receptor TrkA (Leiser et al. 2010), revealing additional potential physiological role(s) of TUFT1. Taken together these studies suggest that TUFT1 could have various functions. However, although various putative functional sites were predicted and the protein sequence was well conserved in the few mammalian species studied (Mao et al. 2001; Deutsch et al. 2002), a connection between TUFT1 structure and its functional sites has not been demonstrated. However, the previous analysis of the TUFT1 sequence was performed more than a decade ago (Deutsch et al. 2002) and it is possible that new functional sites and/or domains could be identified.

The aim of the present study was to identify putative functional positions and domains in the TUFT1 sequence by means of evolutionary analysis of this protein in mammals, i.e. covering circa 200 million years of evolution in this lineage. Such analyses, which can reveal unchanged positions over long geological times and hence strong functional constraints, have proven to be efficient in highlighting important positions. We also investigated putative TUFT1 relationships as they have not been defined.

MATERIALS AND METHODS

Tuftelin sequences and alignment

A total of 82 coding sequences of mammalian *TUFT1* were obtained from *Ensembl* [http://www.ensembl.org] and *NCBI* [http://www.ncbi.nlm.nih.gov] databases. Species names and sequence references are listed in supplementary material 1 (SM1). The dataset was built with: six published full-length sequences (human, orangutan, baboon, cow, mouse and rat); 71 computer-predicted sequences, i.e. available from the automatic analysis of sequenced, or currently being sequenced, mammalian genomes; and five sequences obtained using BLAST from the whole genome shotgun (WGS) repository sequences. *TUFT1* sequences were individually checked through alignment to published cDNA sequences, with particular attention paid to the intron/exon boundaries. When necessary, the sequences were corrected and/or completed using a BLAST search against the WGS sequences. We performed codon alignments by projecting the results of the amino acid alignment onto the nucleotide sequences using Clustal X 2.0 (Higgins et al. 1996). We chosed Clustal X 2.0 because the sequences are conserved, relatively short and contain few gaps. Putative problems of alignment generated by Clustal X 2.0 analyses were also tested using MUSCLE (Edgar, 2004). The same alignment was obtained using both alignment tools. The 82 TUFT1 sequences are provided in the supplementary material 2 (SM2). A single transcript was identified in each species. Our final alignment is available in the supplementary material 3 (SM3).

In our final alignment, only 1,000 residues out of 32,902 were missing, representing 3% of the data. The positions with missing data were included in our analyses and treated as "unknown states". Gaps were removed in our evolutionary computations.

Sliding window and non-synonymous substitution rate (dN) analyses

To identify strong functional constraints, a sliding window analysis and a non-synonymous substitution rate analysis of nucleotide sequence variability were conducted on our alignment using HYPHY (Kosakovsky Pond et al. 2005; Kosakovsky Pond and Muse 2005; Kosakovsky Pond and Frost 2005). HYPHY utilizes Ln likelihood to measure the selective pressure. For every variable site, four quantities were computed: normalized expected numbers (ES and EN) and observed numbers (NS and NN) of synonymous and non-synonymous substitutions, respectively. HYPHY estimates $dN = NN/EN$ and $dS = NS/ES$. The p-value derived from a two-tailed extended binomial distribution is used to assess significance. The test assumes that under neutrality, a random substitution will be synonymous with probability $P = ES/(ES+EN)$, and computes how likely it is that P, NS out of NN+NS substitutions are synonymous. At each position, the probability for the observed data is calculated by the likelihood algorithm, taking into account phylogenetic relationships. In the two analyses, we worked with the "local model", in which all model parameters are estimated independently for each branch, and it was computed using tree topology found in a recent mammalian phylogeny (Meredith et al. 2011).

The non-synonymous substitution rate (dN) was calculated using the maximum likelihood method based on the HKY 85 model (Hasegawa et al. 1985).

The Sliding Window (available in "standard analyses" section and "Miscellaneous" sub-section in HYPHY software) is a method to calculate the mean substitution rate along a protein sequence. This graphical representation is used to visualize selective pressures along the protein. Indeed, nucleotide diversity reflects selective constraints: the lower the variability, the higher the selection, and vice versa.

We chose the following parameters:

- the mean substitution rate was calculated using the maximum likelihood (ML) method based on the HKY 85 model (Hasegawa et al. 1985);

- probabilities are calculated for a window of 15 bp with an overlap of 5 bp between each pair of windows. Other investigators have chosen larger windows for their analyses (e.g. Endo et al. 1996; Tsunoyama and Gojobori 1998; Schmid and Yang 2008) but, when applying the HYPHY method, it is not necessary to use large sliding windows and it is even recommended to avoid large windows, which produce a ''smoothing'' effect that could result in loss of evolutionary information.

Distance tree

The alignment was treated by MEGA 5.2.2 (Tamura et al. 2011) software (http://www.megasoftware.net). The phylogenetic reconstructions used Neighbor-joining and Maximum Likelihood methods, with Dayhoff model and a rate of substitution Gamma distributed (in both NJ and ML).

Purifying selection analysis

The search for site-specific purifying selection (i.e. biologically significant amino acids) in TUFT1 was carried out using the Consurf Server 2.4 (http://consurf.tau.ac.il/) (Ashkenazy et al. 2010; Celniker et al. 2013). The analysis was performed by comparing a null model, i.e. no purifying selection, and a model allowing purifying selections (HKY 85 model: Hasegawa et al. 1985). The results were then displayed on the human sequence. Different levels of purifying selection were indicated by a set of colours.

Positive selection analysis

The search for site-specific positive selection in TUFT1 was carried out using the Selecton Server (http://selecton.tau.ac.il/) (Stern et al. 2007). The Selecton analysis was computed using tree topology in a recent mammalian phylogeny (Meredith et al. 2011). The analysis was performed using the M8 model (Yang et al. 2000). A proportion p0 of the sites was drawn from a beta distribution (which is defined in the interval [0,1]), and a proportion p1(= 1 - p0) of the sites was drawn from an additional category ωs (which is constrained to be ≥ 1). Thus, sites drawn from the beta distribution were sites experiencing purifying selection, whereas sites drawn from the ωs category are sites experiencing either neutral or positive selection. Both p0 and ωs are estimated using ML. The results were then displayed on the human sequence. Different levels of positive selection were indicated by a set of two colours.

Model = Positive selection enabled (M8, beta + w ≥ 1); number of categories = 8.

Putative functional sites

Search for post-translationally modified sites in the human TUFT1 sequence was performed using the Prosite database [Sigrist et al. 2010; http://prosite.expasy.org/] to identify putative N-glycosylation and phosphorylation sites.

Date calibration

Date calibration of the mammalian lineage was obtained from the "Fossil Calibration Database" [http://fossilcalibrations.org]. The mammalian node is dated from 164.9 Ma to 201.5 Ma.

Relationships

To improve our understanding of TUFT1 origins and possible functions, its putative

relationships with other proteins were searched for by using PSI-BLAST (Position-specific

iterative Blast) tool in NCBI site.

RESULTS

Alignment and sequence comparisons

The 82 TUFT1 sequences studied here represent 59 families distributed in 14 orders

(including 1 Monotremata, 3 Marsupiala and 78 Placentalia species) and they are

representative of the current mammalian phylogenetic diversity (SM1). The length of the

TUFT1 sequences ranged from 383 residues in *Choloepus hoffmanii* (sloth) to 399 in

Echinops telfairi (tenrec) with 390 amino acids in humans (SM2). Alignment of these

sequences against the human sequence resulted in a 419 amino acid sequence that included 29

gaps (SM3). Our alignment indicated that TUFT1 is composed of a succession of conserved

and variable regions and revealed only a few insertions of amino acids, the largest being eight

residues inserted in the region encoded by the 3' end of exon 1 in golden mole and five

residues by the 3' end of exon 6 in tenrec (two Afrotheria species). Insertions and deletions of

one to three residues were found in a few sequences and no sequence repeats were identified.

Purifying selection

Analysis of non-synonymous substitution rate (Fig. 1A) and Sliding Window analysis (Fig. 1B)

The dN/dS ratio analysis revealed that TUFT1 sequence is characterized by regions of weak

selective pressures alternating with regions of strong selective pressures (Fig. 1A, B). The

regions subjected to strong functional constraints display low -Ln likelihood and low dN

values. The lower these values are, the more important the selective constraint was (for this

analysis the limits were determined arbitrarily; for example -Ln Likelihood < 250-200 and dN < 0.6). The strongest constraints were found in the regions/residues encoded by exon 3, exon 4, the 5' end of exon 5, and exons 10 to 12. In contrast, the regions encoded by exons 6 to 9 were found to be less conserved, meaning they were less subjected to functional constraints.

Distance tree

The phylogenetic reconstruction was performed to (i) determine the rate of TUFT1 evolution and (ii) check anomalies (long branches) occurring during evolution in some mammalian lineages to avoid using a sequence that is no longer under selective pressure (i.e. a pseudogene). Indeed, if TUFT1 sequences were conserved during mammalian evolution, phylogenetic information would still be present and we would expect to obtain a phylogenetic tree similar to that currently considered for mammalian relationships; otherwise, we would expect to obtain long branches along with incorrect topology. We present an NJ tree in figure 1, but the ML tree is available in SM9.

The distance tree indicates a good equilibrium between conserved and variable regions of TUFT1 sequences (Fig. 2).

A few *TUFT1* sequences display long branches, a feature which means a more rapid accumulation of substitutions without changing the phylogenetic relations. This is the case for manatee (*Trichechus manatus*) and tenrec (*Echinops telfairi*) in Afrotheria, bushbaby (*Otolemur garnetii*) in Primates, guinea pig (*Cavia porcellus*) in Rodentia, hedgehog (*Erinaceus europaeus*) in Laurasiatheria, and opossum (*Monodelphis domestica*) in Marsupiala. Very few sequences are not located at the correct phylogenetic position such as the shrew (*Sorex araneus*) and the kangaroo rat *Dipodomys ordii*. However, the *TUFT1* sequence of *Dipodomys ordii* was not complete enough to have a good signal (SM2).

Consurf analysis

The statistical tests supporting this analysis are available in SM6.

Purifying selection detected by Selecton analysis at the amino acid level identified TUFT1 residues that were well conserved during mammalian evolution, i.e. that are predicted to play an important functional or structural role, or both (Fig. 3, SM4). A total of 121 conserved (= unchanged) and 47 conservative (= that could be only replaced with a residue possessing similar properties) positions were identified. These important positions represent more than 43% of the 390 residues composing the human TUFT1 sequence. These residues are either isolated or regrouped, forming motifs of various lengths:

(1) At the N-terminus, the motif ^{1}MNGT contains a putative N-glycosylation site (^{2}N).

(2) The N-terminal region includes a large motif encoded by the 3' region of exon 3 to the 5' region of exon 5. It is composed of 57 residues, starting at ^{64}S and ending at ^{121}S. This motif contains 33 unchanged or conservative positions. The function of this motif is unknown to date, with the exception of the last residue ^{121}S, which is putatively phosphorylated (protein kinase C site).

(3) The C-terminal region contains another large motif, encoded by exons 10 to 12, consisting of 74 residues, from ^{278}E to ^{351}Q. This motif possesses 55 unchanged or conservative positions. Most of this motif is known to belong to the TIP-39 binding site.

(4) At the C-terminal extremity encoded by exon 13, the ^{381}PmPvIRVVET motif of unknown function is well conserved.

In addition to these motifs, our evolutionary analysis highlighted many conserved or conservative residues, hence those having a putative function. Among them, some were predicted as functional by *Prosite* analysis in the human sequence (see below, Fig. 4).

Positive selection

Using the Selecton server, 13 positions were detected as positively selected during mammalian evolution, i.e. sites drawn from the ωs category (Stern et al., 2007). Shades of

yellow indicate ω>1, with dark-yellow representing sites where reliable positive selection was inferred, and light-yellow representing positive selection that was not statistically significant (SM4, SM8). These 13 positions were indicated on the human sequence (Fig. 3). One is isolated, at [255]V, and 12 are regrouped, forming three motifs under positive selection: [131]SLHR; [153]AlYssPP; [163]TCI. One of them contained a "casein kinase II" phosphorylation site predicted by *Prosite* on the human sequence (see below).

Prediction of post translation modifications

Prosite could not predict any functional sites, however, it predicted 14 sites that undergo post-translation modifications, and hence these positions may be involved in TUFT1 function (SM5). Most of these sites were also predicted in a previous analysis (Deutsch et al. 2002). These were: two N-glycosylation sites (positions 2 and 356), six casein kinase II (CK2) phosphorylation sites (phosphoserines 122, 157 and phosphothreonines 9, 35, 175, 322), and four protein kinase (PKC) phosphorylation sites (phosphoserines 121, 171, 370, 378). One amidation site (46) and one N-myristoylation site (355), close to the N-gly 356, were also predicted. All the *Prosite*-predicted sites were reported on the 3D TUFT1 sequence published by Deutsch et al. (2002) including sites that were considered as functional in our evolutionary analysis (Fig. 4). Cysteines were also indicated in this figure.

Relationships of TUFT1

Using PSI-BLAST with the human TUFT1, we found a single related protein, MYZAP (myocardial zonula adherens protein), that is known in many vertebrate taxa (including coelacanth and teleost fishes). This finding strongly suggests that MYZAP and TUFT1 were present in the last common ancestor of osteichthyans. In humans, *MYZAP* is located on chromosome 15 (*vs* chr 1 for *TUFT1*), is composed of 13 exons, and encodes a cytoplasmic

protein that does not possess a signal peptide, like TUFT1. MYZAP is expressed in adherens junctions of myocardiac and vascular endothelia cells, and in junctions of various epithelia (Rickelt et al. 2011; Pieperhoff et al. 2012). The alignment of TUFT1 and MYZAP points to only a few similarities.

DISCUSSION

Purifying selection

During millions years of evolution, purifying selection does not retain mutations that lead to deleterious effects. Such selection results in the preservation of all sensitive positions, i.e. those possessing amino acids that play either functionally or structurally important roles. Residues are either conserved or conservative. In the TUFT1 sequences, our evolutionary analysis identified a total of 168 conserved and conservative positions during 164.9 Ma of mammalian evolution indicating that the important functions supported by these positions were present earlier in TUFT1 evolution, in non-mammalian vertebrates. We compare the putative functional sites predicted by *Prosite* with the crucial positions revealed by our evolutionary analysis, then we discuss the presence of various conserved domains (Figs 3, 4, SM5).

Several predicted functional sites are confirmed by evolutionary analysis

N-glycosylated residues. - The two predicted N-linked glycosylation sites (N-gly 2, ^{2}**NGT**R and N-gly 356,^{356}**NFST**) are under purifying selection, which strongly suggests that these post-translational modifications (attachment of a glycan, N-acetylglucosamine, to asparagine) of TUFT1are functionally and/or structurally important.

N-myristoylation site. - The predicted N-myristoylation site (N-Myr 355, ^{355}**GNfsTQ**), located in the C-terminal region, near the second N-Gly conserved site, was not validated by our

analysis. Indeed, the glycine (G), which is the crucial, post-translationally modified residue of such a site in the N-terminal region, is located at a variable position.

Phosphorylated sites. - Our analysis indicated that only three out of the ten putative phosphorylated residues predicted by Prosite are under purifying selection. Only phosphorylation of serines by protein kinase C (PKC) were validated by our analysis, through conservation of the position housing a serine: Ser-P 121, [121]**S**sK; Ser-P 171, [171]**S**lR; Ser-P 378, [378]**S**pK. Previous post-translational modification analyses using MS/MS sequencing of the recombinant human TUFT1 protein produced in a eukaryotic system, also revealed one phosphorylated site at Serine 378 (Shay et al., 2009).

None of the casein kinase II phosphorylation sites were confirmed. In a previous study, Deutsch et al. (2002) suggested that phosphorylation sites might play an important role in providing potential sites for specific chelation of calcium ions, which could explain why TUFT1 is presumed to play an important role in enamel mineralization. Our findings could support such a role, although several, previously predicted phosphorylation sites were not conserved during mammalian evolution.

Amidation site. - The amidation site (amidation 46, [46]aGRK) predicted by Prosite is not under purifying selection since all residues of this motif, including the glycine residue that provides the amine group, are located at variable positions according to our analysis.

Evaluation of previously identified functional domains by evolutionary analysis

TUFT1-TIP39 interaction domain. - More than 15 years ago, Paine et al. (2000) showed that TUFT1 interacts with TIP39 (Tuftelin Interacting Protein – 39 kDa; also known as TFIP11 - Tuftelin Interacting Protein 11). These authors showed that the region of interaction with TIP39 is located near the TUFT1 C-terminus, between amino acids 294 and 348 (Paine et al. 2000). This part of the protein was also suggested to be responsible for TUFT1 self-assembly (Paine et al. 2000). Our analysis confirmed the importance of this TUFT1 region because 42

out of the 54 amino acids putatively implicated in this interaction were conserved. Only 12 amino acids are at a variable position. In contrast, most of the conserved and conservative residues are clustered in long stretches within this region, a finding that indicates a crucial role. In addition, we also found that several conserved residues are located at both extremities of this domain (residues 285-293 and 349-360), which indicates that the functional region is probably wider than the region previously identified by Paine et al. (2000). TIP39 (TFIP11) is a nuclear speckle-localized protein that may play a role in spliceosome disassembly in Cajal bodies (Wen et al. 2005; Stanek et al. 2008). In adult tissues, TUFT1 is expressed mainly in the cytoplasm. However, nuclear localization of the protein was demonstrated in several tissues (Deutsch et al. 2002; Leiser et al. 2007). During early mouse embryonic development, TUFT1 was detected mainly in the cytoplasm, while at later embryonic stages and post-natally, its expression in neuronal cells is concentrated in the perinuclear/nuclear region (Deutsch unpublished; Shilo et al. unpublished), possibly indicating interaction with TIP39. *Calcium-binding domain*. - An EF-hand, calcium-binding domain was previously suggested in human, bovine, and murine TUFT1, from residue 125 to 137 (Mao et al. 2001; Deutsch et al. 2002). However, no experimental results supported this motif. Our analysis shows that only one of the residues in this region is conserved in mammals, a finding which indicates that this region lacks an important function (but see below, *Positive selection*).

Domains of unknown function

Our evolutionary analysis predicts that four additional motifs could play an important role not previously identified, either through protein analysis software or in other previous studies. Indeed, these positions are subjected to strong purifying selection. Three are located between residues 64 and 121, whereas the fourth constitutes the C-terminus of TUFT1. Each of them possesses one or two glutamic acid (E) and one lysine (K) or arginine (R). Remarkable positions highlighted by our evolutionary analyses are summarized in SM7.

Positive selection

Positive selection means that one allele was selected because it improves fitness. Therefore, positive selection increases the prevalence of adaptive traits. The results of purifying selection are generally easy to interpret. However, this is not the case for the results of positive selection, because the residues detected as being positively selected are often not related to a motif, and hence interpretation of a putative role is difficult. Another problem of positive selection is the possibility of obtaining false positive results. However, our analysis of TUFT1 indicates that several positively-selected positions are located in close proximity to one another, and belong to two putative motifs, which were not under purifying selection.

Four positions under positive selection ([131]SLHR) are located in a TUFT1 region that was previously described as a "calcium-binding domain", but predicted to be variable in our evolutionary analysis. We do not know the role of this motif, however, its unknown function was acquired recently in mammalian history, most probably in placental mammals, which could explain why that motif was not detected as being important in our analysis. Therefore, the potential role of TUFT1 in mineralisation (Deutsch et al. 2002) could be a recent feature, on the geological scale.

The second motif revealed as being positively selected is the [157]sPPe encoded by the 3' region of exon 6. This short sequence was predicted by Prosite as a casein kinase II phosphorylation site (SppE) but not validated by our evolutionary analysis of conserved positions. As discussed above, the positive selection of the two prolines probably occurred recently in the mammalian lineage and could mean that this motif is functional in placental mammals for example, strengthening the phosphorylation pattern of TUFT1 (Deutsch et al. 2002). A mutated enamelin phosphorylation site causes amelogenesis imperfecta (Chan et al. 2010), which emphasizes the importance of conservation of phosphorylation pattern during evolution of secretory calcium-binding phosphoproteins (Al-Hashimi et al. 2009; Silvent et al.

2013). Our results confirmed the importance of mineralization among the different functions of these ubiquitous proteins by reinforcement of phosphorylation.

In addition, the presence of a positively selected valine at position 255, close to the well-conserved [256]ALEE motif, could increase the fitness and the role of this highly conserved motif, although its role is not yet known.

Finally, the putative role of the [163]TCI motif encoded by exon 7 and composed of three positively selected residues is unknown. The presence of a cysteine could be important if it were involved in the formation of a disulfide bond that could promote a new tertiary structure for TUFT1. However, none of the other cysteines in TUFT1 are conserved, which does not support such a model.

Alternative splicing

Alternatively spliced TUFT1 mRNA transcripts have been detected in different tissues (Mao et al. 2001; Deutsch et al. 2002). An isoform lacking exon 2 was identified in the mouse kidney, whereas another lacking exon 2, part of exon 4, exon 5-11 and part of exon 12 was found in mouse liver. Various TUFT1 isoforms were also identified in tooth buds (lacking exon 2, or 3 or 6). Both the lack of exons 5-9 in some transcripts and the fact that they are variable could indicate their recent recruitment, i.e. in tetrapod ancestors for example. This hypothesis could also explain the presence of amino acids under positive selection in exons 6 and 7 (new constraints resulting from new selective pressures). It should be noted that these two alternatively spliced exons each have a two-amino acid motif under strong positive selection.

Relationships of TUFT1

The single protein we found related to TUFT1 is MYZAP. Both proteins share some amino acid sequence similarities and lack a signal peptide, and their genes display similar exonic structures. These findings suggest that one might have derived from the other after duplication of an ancestral gene. Their probable presence in the genome of ancestral osteichthyans and their location on different chromosomes could indicate that their origin dates back to the genome duplication that occurred at the onset of vertebrate diversification, approximately 500 million years ago. The long history of these two proteins could explain why they do not share many similarities.

Additional remarks

Recent functional studies pointed to the involvement of TUFT1 in adaptation to hypoxia and in differentiation of neurons (Leiser et al. 2010, Deutsch et al. 2011). However, no interacting proteins or sequences within TUFT1 were indicated as being involved in these functions. Two putative HIF consensus DNA binding sites, compatible with hypoxia responsive elements (HRE, 5'-RCGTG-3') in the *TUFT1* promoter region, at positions -1296 and -27 of human *TUFT1* (upstream to exon 1; GenBank: AH009496.1), and at positions -130 and -8 of the mouse *TUFT1* promoter (GenBank: NC_000069.5), were identified (Leiser et al. 2010). These sequences, which were not included in the current analyses, point to the regulatory role of HIF1a on TUFT1 expression.

TUFT1 is an acidic protein and although predictions of its secondary structure reveals two long coiled-coil regions, consurf and other prediction algorithms indicate that most residues are exposed, whereas a few buried residues could contribute to 3D structure. Hence, analysis of linear motifs was performed to reveal motifs predicting function. Unlike amelogenin, a search for linear motifs using the eukaryotic linear motif (ELM) server (Dinkel et al. 2016) revealed no motifs. The major contribution of our consurf analysis was to

highlight four unknown regions under strong purifying selection and indicate functional and

buried residues, while previous studies can only point out (i) motifs indicated by yeast two-

hybrid systems (TIP39 and self-assembly), (ii) results of previous *in silico* analysis (Ca

binding domain), or (iii) post-translational modifications that might point to a function.

References

Al-Hashimi N, Sire JY, Delgado S (2009) Evolutionary Analysis of Mammalian Enamelin, the Largest Enamel Protein, Supports a Crucial Role for the 32 kDa Peptide and Reveals Selective Adaptation in Rodents and Primates. J Mol Evol 69(6):635-656.

Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res 38 (Web Server issue):W529-533.

Bashir MM, Abrams WR, Tucker T, Sellinger B, Budarf M, Emanuel B, Rosenbloom J (1998) Molecular cloning and characterization of the bovine and human tuftelin genes. Connect. Tissue Res 39:13-24.

Celniker G, Nimrod G, Ashkenazy H, Glaser F, Martz E, Mayrose I, Pupko T, Ben-Tal N (2013) ConSurf: Using evolutionary data to raise testable hypotheses about protein function. Isr J Chem 53:199-206.

Chan HC, Mai L, Oikonomopoulou A, Chan HL, Richardson AS, Wang SK, Simmer JP, Hu JC (2010) Altered enamelin phosphorylation site causes amelogenesis imperfecta. J Dent Res 89:695-699.

Deutsch D (1989) Structure and function of enamel gene products. Anat Rec 224:189-210.

Deutsch D, Palmon A, Fisher LW, Kolodny N, Termine JD, Young MF (1991) Sequencing of bovine enamelin ("tuftelin") a novel acidic enamel protein. J Biol Chem 266:16021-16028.

Deutsch D, Palmon A, Young MF, Selig S, Kearns WG, Fisher LW (1994) Mapping of the human tuftelin (TUFT1) gene to chromosome 1 by fluorescence in situ hybridization. Mamm. Genome 5:461-462.

Deutsch D, Palmon A, Dafni L, Catalano-Sherman J, Young MF, Fisher LW (1995) The enamelin (tuftelin) gene. Int Dev Biol 39:135-143.

Deutsch D, Palmon A, Dafni L, Mao Z, Leytin V, Young M, Fisher LW (1998) Tuftelin – aspects of protein and gene structure. Eur J Oral Sci 106 (Suppl. 1): 315–323.

Deutsch D, Shay B, Rosenfeld E, Leiser Y, Fermon E, Taylor A, Charuvi K, Cohen Y, Haze A, Fuks A, Dafni L, Mao Z (2002) The human tuftelin gene and the expression of tuftelin in mineralizing and nonmineralizing tissues. Connect Tissue Res 43:425-434.

Deutsch D, Silverstein N, Shilo D, Lecht S, Lazarovici P, Blumenfeld A (2011) Biphasic influence of hypoxia on tuftelin expression in mouse mesenchymal C3H10T1/2 stem cells. Eur J Oral Sci 119 (suppl.1):55-61.

Dinkel H, Van Roey K, Michael S, Kumar M, Uyar B, Altenberg B, Milchevskaya V, Schneider M, Kühn H, Behrendt A, Dahl SL, Damerell V, Diebel S, Kalman S, Klein S,

Knudsen AC, Mäder C, Merrill S, Staudt A, Thiel V, Welti L, Davey NE, Diella F, Gibson TJ (2016) ELM 2016-data update and new functionality of the eukaryotic linear motif resource. Nucleic Acids Res 44(D1):D294-300.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5):1792-1797.

Endo T, Ikeo K, Gojobori T (1996) Large-scale search for genes on which positive selection may operate. Mol Biol Evol 13:685-690.

Jeremias FL, Koruyucu M, Küchler EC, Bayram M, Tuna EB, Deeley K, Pierri RA, Souza JF, Fragelli CM, Paschoal MA, Gencay K, Seymen F, Caminaga RM, dos Santos-Pinto L, Vieira AR (2013) Genes expressed in dental enamel development are associated with molar-incisor hypomineralization. Arch Oral Biol 58:1434-1442.

Kosakovsky Pond SL, Frost SDW (2005) A genetic algorithm approach to detecting lineage-specific variation in selection pressure. Mol Biol Evol 22:478-485.

Kosakovsky Pond SL, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676-679.

Leiser Y, Blumenfeld A, Haze A, Dafni L, Taylor AL, Rosenfeld E, Fermon, E, Gruenbaum-Cohen Y, Shay B, Deutsch D (2007) Localization, quantification, and characterization of tuftelin in soft tissues. Anat Rec 290:449-454.

Leiser Y, Silverstein NC, Blumenfeld A, Shilo D, Haze A, Rosenfeld E, Shay B, Tabakman R, Lecht S, Lazarovici P, Deutsch D (2010) The induction of tuftelin expression in PC12 cell line during hypoxia and NGF induced differentiation. J. Cell Physiol 226:165-172.

MacDougall M, Simmons D, Dodds A, Knight C, Luan X, Zeichner-David M, Zhang C, Ryu OH, Qian Q, Simmer JP, Hu C-C (1998) Cloning, characterization, and tissue expression pattern of mouse tuftelin cDNA. J Dent Res 77:1970-1978.

Mao Z, Shay B, Hekmati M, Fermon E, Taylor A, Dafni L, Heikenheimo K, Lustmann J, Fisher LW, Young MF, Deutsch D (2001) The human tuftelin gene: cloning and characterization. Gene 279:181-196.

Meredith RW1, Janečka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simão TL, Stadler T, Rabosky DL, Honeycutt RL, Flynn JJ, Ingram CM, Steiner C, Williams TL, Robinson TJ, Burk-Herrick A, Westerman M, Ayoub NA, Springer MS, Murphy WJ (2011) Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. Science 334(6055):521-524.

Paine CT, Paine ML, Luo W, Okamoto CT, Lyngstadaas SP, Snead ML (2000) A tuftelin-interacting protein (TIP39) localizes tothe apical secretory pole of mouse ameloblasts. J Biol Chem 275:22284-22292.

Patir A, Seymen F, Yildirim M, Deeley K, Cooper ME, Marazita ML, Vieira AR (2008) Enamel formation genes are associated with high caries experience in Turkish children. Caries Res 42:394-400.

Pieperhoff S, Rickelt S, Heid H, Claycomb WC, Zimbelmann R, Kuhn C, Winter-Simanowski S, Kuhn C, Frey N, Franke WW (2012) The plaque protein myozap identified as a novel major component of adhering junctions in endothelia of the blood and the lymph vascular systems.J Cell Mol Med 16:1709-1719.

Rickelt S, Kuhn C, Winter-Simanowski S, Zimbelmann R, Frey N, Franke WW (2011) Protein myozap--a late addition to the molecular ensembles of various kinds of adherens

junctions.Cell Tissue Res 346:347-359.

Schmid K, Yang Z (2008) The trouble with sliding windows and the selective pressure in BRCA1. PLoS ONE 3:e3746

Sergei L. Kosakovsky Pond SL, Spencer V, Muse SV (2005) HyPhy: Hypothesis testing using Phylogenies. In Statistical Methods for Molecular Evolution. Statistics for Biology and Health, Part II, pp. 125-181.

Shay B, Gruenbaum-Cohen Y, Tucker AS, Taylor AL, Rosenfeld E, Haze A, Dafni L, Leiser Y, Fermon E, Danieli T, Blumenfeld A, Deutsch D (2009) High yield expression of biologically active recombinant full length human tuftelin protein in baculovirus-infected insect cells. Protein Expr Purif 68:90-98.

Silvent J, Sire JY, Delgado S (2013) The Dentin Matrix Acidic Phosphoprotein 1 (DMP1) in the light of mammalian evolution. J Mol Evol 76(1-2):59-70.

Stanek, D, Pridalova-Hnilicova J, Novotny I, Huranova M, Blazikova M, Wen X, Sapra AK, Neugebauer KM (2008) Spliceosomal small nuclear ribonucleoprotein particles repeatedly cycle through Cajal bodies. Mol Biol Cell 19:2534-2543.

Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E, Pupko T (2007) Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Res 35(Web Server issue):W506-511.

Sigrist CJA, Cerutti L, de Castro E, Langendijk-GenevauxPS, Bulliard V, Bairoch A, Hulo N (2010) PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res 38(Database issue):D161-D166.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731-2739.

Tsunoyama K, Gojobori T (1998) Evolution of nicotinic acetylcholine receptor subunits. Mol Biol Evol 15:518-527.

Wen X, Lei Y-P, Zhou YL, Okamoto CT, Snead ML, Paine ML (2005) Structural organization and cellular localization of tuftelin-interacting protein 11 (TFIP11). Cell Mol Life Sci 62:1038-1046.

Figure Captions

Figure 1. Evolutionary analysis of the mammalian TUFT1 sequences using SLAC (A) and

the Sliding Window (B) analysis.

A. Non-synonymous substitution rate (dN) along TUFT1 sequences.

B. Logarithm of substitution rate per site along the TUFT1 sequences estimated for a window

of 15 bp with an overlap of 5 bp between each pair of windows.

The regions with the lowest rate of non-synonymous substitution (A) and the lower Ln

likelihood (B) indicate strong constraints, which reflect high selective pressures.

Figure 2. Distance tree obtained from alignment of the 82 mammalian *TUFT1* sequences.

The longer the branches are, the higher the evolutionary rate of the taxa. Scale bar = number

of substitutions per site.

Figure 3. Evolutionary chart of TUFT1 calculated from the 82 mammalian sequences. This

chart is a simplification of Consurf analysis (SM4), and was deduced from the results

obtained when dN/dS was calculated at each codon of TUFT1 by Conseq (purifying

selection) and by Selecton (positive selection). The human sequence was used as reference.

Positions subjected to purifying selection are marked as black (conserved positions) and grey

(conservative positions) background. Positively selected positions are indicated with an

asterisk. The two frames represent putative functional domains predicted in previous studies

(see references in text).

Figure 4. Schematic representation of the human TUFT1 sequence, on which are indicated

the domains and remarkable positions identified in previous studies and present work (after

Deutsch et al., 2002). The regions and amino acid positions that were conserved during 200

million years of mammalian evolution are indicated in green. They are predicted to play an important role either for the function or the structure of the protein. In contrast, the regions and positions that are considered variable are indicated in red. It is worth noting that all cysteines and most positions predicted by *Prosite* as being functional were not validated by our analyses.

Supplementary material 1 (SM1). Scientific names, common names, families, orders and references in GenBank for the 82 *TUFT1* sequences used in our study. Published sequences in bold. Alphabetical order of Latin names. Amino acid sequences are available in SM2. XM_ and NM_ sequences were obtained from *NCBI* database. ENS sequences from *Ensembl* release 80 and 81. Other sequences were obtained in GenBank using blast on genomes being sequenced [https://www.ncbi.nlm.nih.gov/Traces/wgs/].

Supplementary material 2 (SM2). Amino acid sequences of the 82 mammalian TUFT1 proteins included in the study. Number of amino acids between brackets. ? = unknown residues.

Supplementary material 3 (SM3). The 82 mammalian TUFT1 sequences were aligned against the human sequence and were ordered following mammalian relationships. (|)= exon limits; (.)= residue identical to human TUFT1 residue; (-)= indel; (?)= unknown amino acid; (*)= stop codon. See SM2 for amino acid sequences.

Supplementary material 4 (SM4). Results of the Consurf analysis of the TUFT1 sequences from 82 mammalian species, revealing a panorama of the various selections acting on TUFT1.

Supplementary material 5 (SM5). Prediction of post translational modifications of the

human TUFT1 sequence using *Prosite* database.

Supplementary material 6 (SM6). Statistical tests supporting Consurf analysis.

Supplementary material 7 (SM7). Non-exhaustive list of remarkable amino acid

positions of TUFT1 highlighted in our study.

Supplementary material 8 (SM8). Results of the Conseq analysis of TUFT1 sequences from

82 mammalian species.
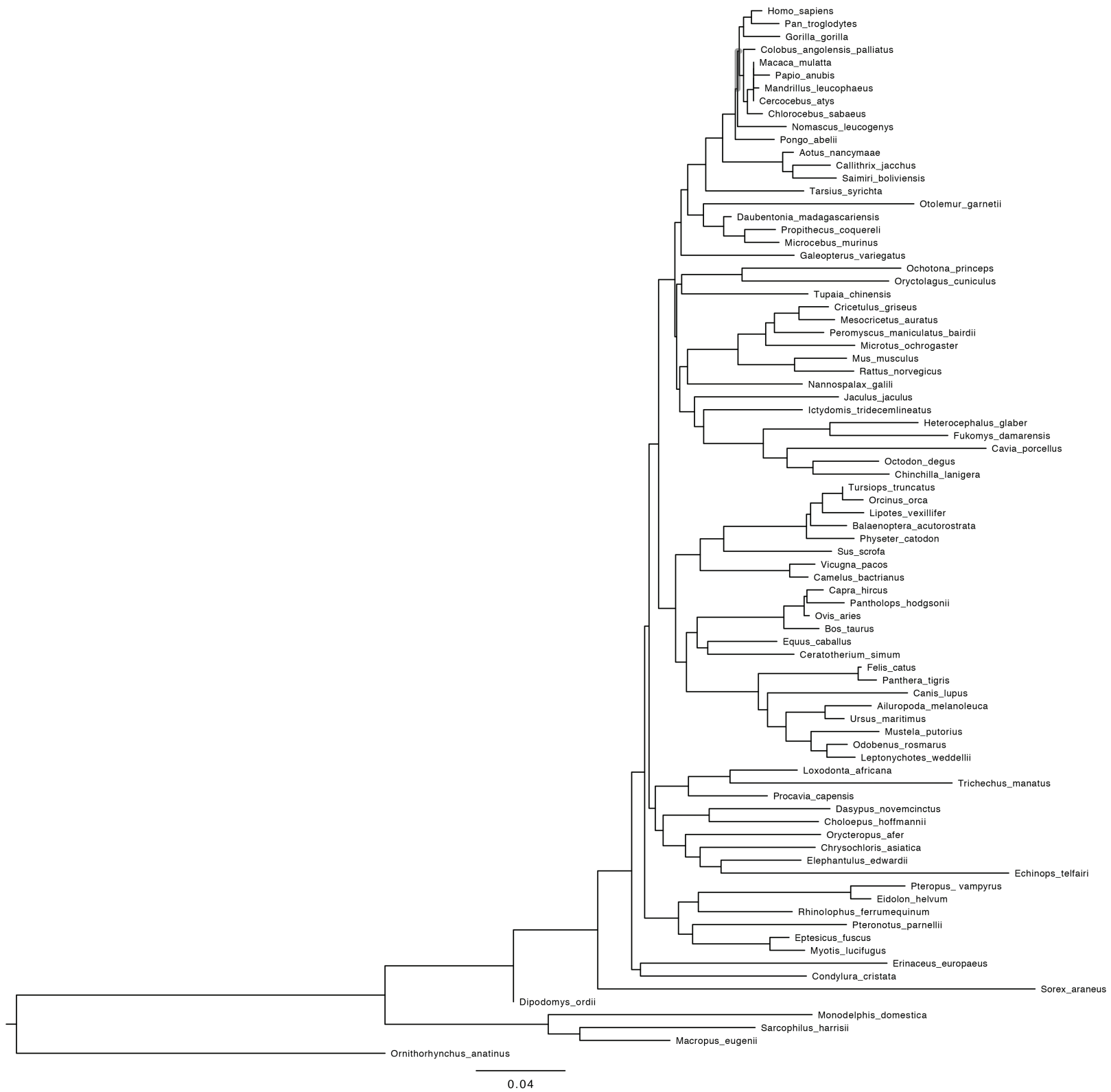
Figure 2. Delgado et al.
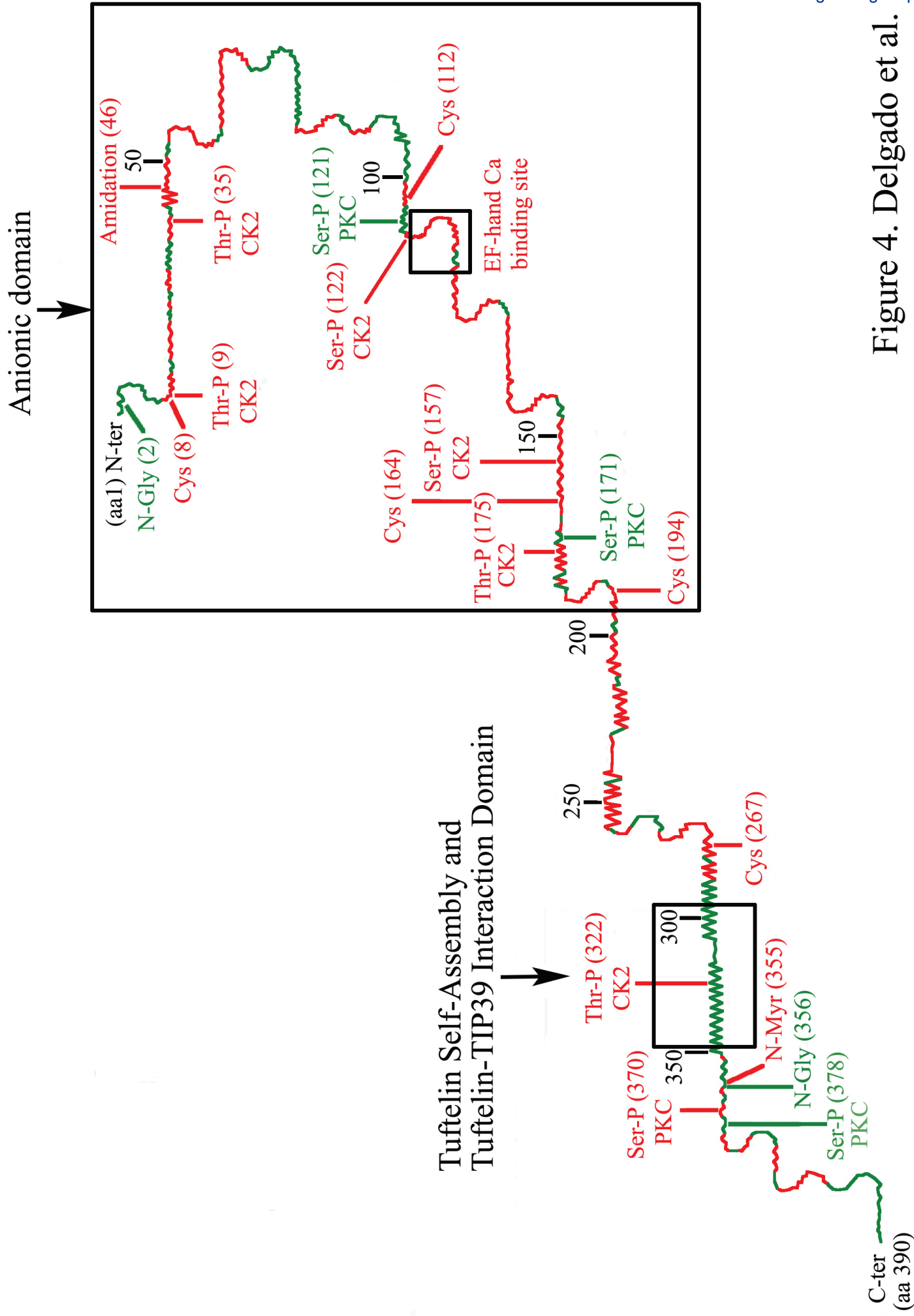
Figure

Figure 4. Delgado et al.

Figure 1 revised

Click here to download Figure Fig 1.eps
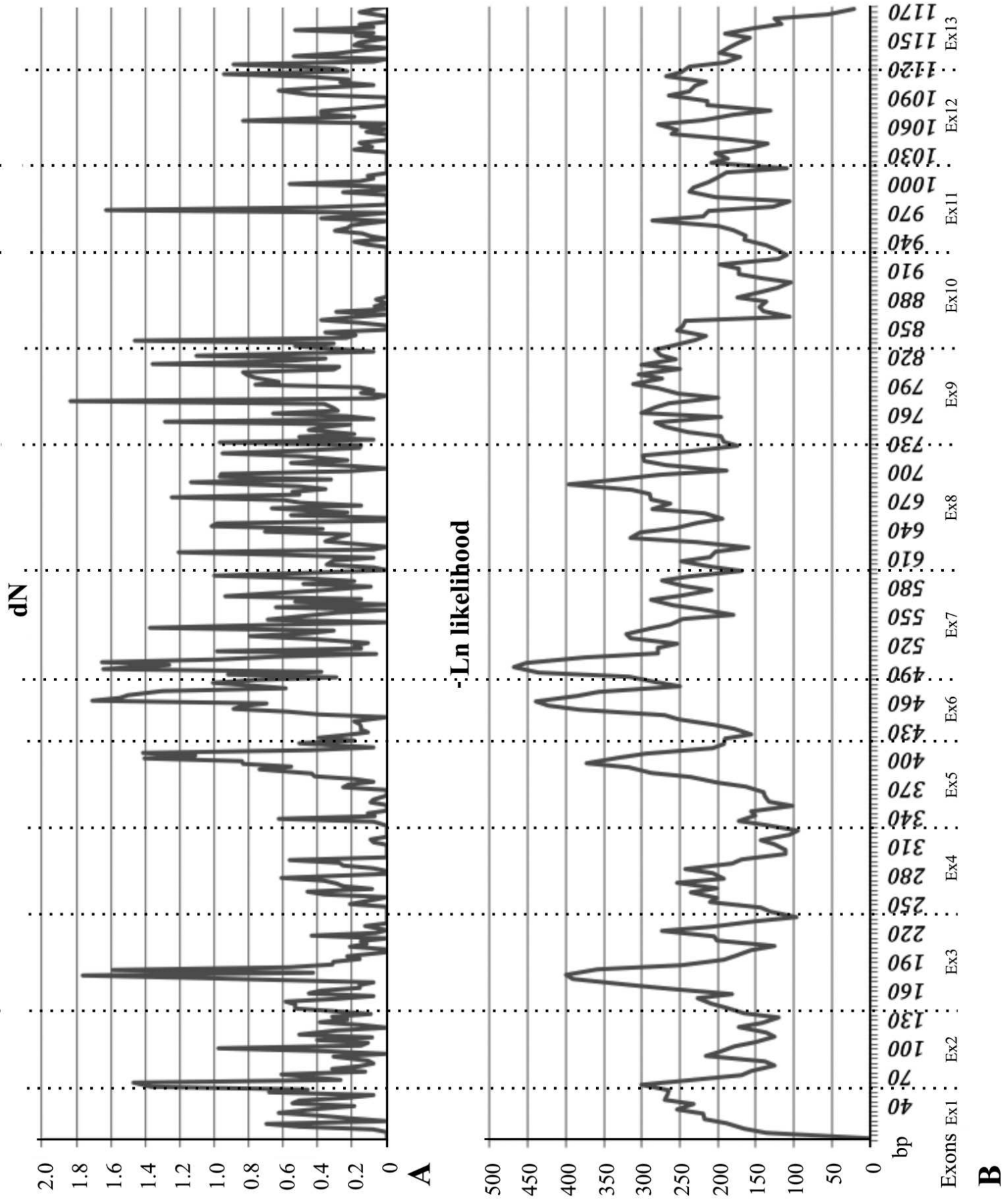
Figure 1. Delgado et al.
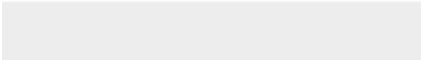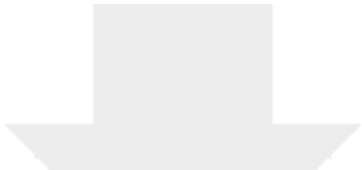
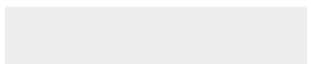Click here to access/download
Supplementary Material
SM9.pdf

Click here to access/download
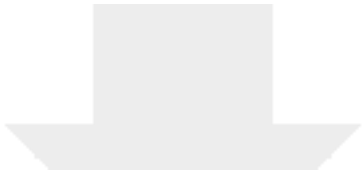**Supplementary Material**
SM2.pdf

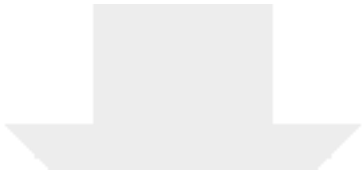Click here to access/download
Supplementary Material
SM3.pdf

Click here to access/download
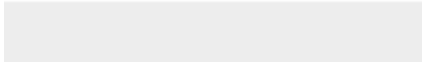**Supplementary Material**
SM4.jpg

Click here to access/download
Supplementary Material
SM5.pdf

Click here to access/download
Supplementary Material
SM1.pdf

Click here to access/download
Supplementary Material
SM6.pdf

Click here to access/download
Supplementary Material
SM7.pdf

Click here to access/download
Supplementary Material
SM8.pdf