THE PROTEIN SOCIETY

Superdomains in the protein structure hierarchy: The case of PTP-C2

Donald T. Haynie[1]* and Bin Xue[2]

[1]Department of Physics, School of Natural Sciences and Mathematics, College of Arts and Sciences, University of South Florida, Tampa, Florida 33620

[2]Department of Cell Biology, Microbiology, and Molecular Biology, School of Natural Sciences and Mathematics, College of Arts and Sciences, University of South Florida, Tampa, Florida 33620

Received 22 December 2014; Accepted 12 February 2015
DOI: 10.1002/pro.2664
Published online 18 February 2015 proteinscience.org

Abstract: *Superdomain* is uniquely defined in this work as a conserved combination of different globular domains in different proteins. The amino acid sequences of 25 structurally and functionally diverse proteins from fungi, plants, and animals have been analyzed in a test of the superdomain hypothesis. Each of the proteins contains a protein tyrosine phosphatase (PTP) domain followed by a C2 domain. Four novel conserved sequence motifs have been identified, one in the PTP domain and three in the C2 domain. All contribute to the PTP-C2 domain interface in PTEN, a tumor suppressor, and all are more conserved than the PTP signature motif, $HCX_3(K/R)XR$, in the 25 sequences. We show that PTP-C2 was formed prior to the fungi, plant, and animal kingdom divergence. A superdomain as defined here does not fit the usual protein structure classification system. The demonstrated existence of one superdomain suggests the existence of others.

Keywords: domain; evolution; hierarchy; protein; structure; superdomain

Introduction

The canonical levels of protein structure— primary, secondary, and tertiary—were defined by Linderstrøm-Lang in 1951.[1] Soon thereafter, the interactions of separate chains in a folded protein came to be called quaternary structure. (See Supporting Information.) More recent research has revealed the structural relatedness of different proteins, for instance, myoglobin and hemoglobin[2] and lysozyme and α-lactalbumin,[3] and the existence of protein superfamilies and domain superfolds.[4] A superfamily is the largest clade, or grouping, for which common ancestry can be inferred, usually by comparison of primary structures, or amino acid sequences. A superfold is one that is adopted by proteins of no apparent structural or functional similarity. Two examples of superfamilies are the protein tyrosine phosphatase (PTP) domain[5] and the protein kinase C2 domain.[6] An example of a superfold is the triosephosphate isomerase barrel, which features both α helices and β strands.[7]

Multi-domain proteins comprise some number of nominally independent folding units, or domains, typically globular in character. Such proteins were unknown to Linderstrøm-Lang. In principle, the individual domains could be encoded by different genes and synthesized as separate polypeptides. Some 4/5 of metazoan proteins have at least two domains,[8] however, suggesting that tethering may confer a fitness advantage or support the formation of more complex structures. An example of a multi-domain protein is fibronectin, which consists of numerous closely related modular units in tandem.[9] Another example is tensin 1 (TNS1), which comprises a PTP domain[5] and a C2 domain[6] near the N-terminus, a Src homology 2 (SH2) domain,[10] and a protein tyrosine binding (PTB) domain[11] near the C-

Abbreviations: C2, protein kinase C2; PTP, protein tyrosine phosphatase.

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Donald T. Haynie, Department of Physics, School of Natural Sciences and Mathematics, College of Arts and Sciences, University of South Florida, Tampa, FL 33620. E-mail: dhaynie@usf.edu

Published by Wiley-Blackwell. © 2015 The Protein Society

terminus, and a large, possibly intrinsically disordered, region in between.[12]

It is possible that Linderstrøm-Lang did not identify all levels of organization relevant to multidomain protein structure. One such level could potentially occur between tertiary structure and quaternary structure. The present study concerns this possibility. We focused on proteins that comprise a PTP domain and C2 domain in a specific orientation in the polypeptide chain; these domains are not known to occur in a different orientation in the same protein. The study was aided by crystallographic information on the well known human tumor suppressor phosphatase homolog/tensin homolog (PTEN), which consists of a PTP domain, a C2 domain and a short linker in-between.[13] The linker consists of HLDYRPV; the tyrosine side chain contributes to the domain interface.

Found in many different proteins, from prokaryotes to humans, PTP domains were identified by associating tyrosine phosphatase activity and amino acid sequence data.[5] At first, PTPs seemed distinct from dual-specificity phosphatases (DSPs), which recognize not only phosphorylated tyrosine but also phosphoserine and phosphothreonine (pThr).[14] Some investigators speculated that the common phosphatase signature motif of PTPs and DSPs, HCX_3(K/R)XR, might be evidence of convergent evolution, as alignment was often poor elsewhere in the chain. The crystal structure of a vaccinia-related human DSP provided crucial evidence for the similarity of its fold to those of *Yersinia pestis* PTP and human PTP1B,[15] but it did not exclude the convergence hypothesis. More recently, TNS1 and PTEN have been identified as PTPs. TNS1 is an adhesion plaque protein in which asparagine substitutes for the essential cysteine nucleophile of the signature motif;[16] PTEN, a DSP, is encoded by a gene that is mutated in numerous cancers.[17,18] PTEN substrates include not only proteins phosphorylated on amino acid side chains but also phosphoinositides.[19] The PTEN crystallographic structure shows that this PTP is a relative of the *Yersinia* PTP, PTP1B and the vaccinia-related human DSP,[13,15] decreasing the plausibility of the convergence hypothesis.

The C2 domain too occurs in many different proteins.[6] Examples include protein kinase C and synaptotagmin I. C2 domains participate in membrane targeting and bind up to two Ca^{2+} ions in some cases.[6] At least one C2 domain binds a pTyr-containing peptide.[20] In PTEN, residues between β strands 11 and 12 may be intrinsically disorganized.[13] Numerous clades of C2 structure have been distinguished.[21] Here, we argue that the PTP and C2 domains of PTEN constitute a superdomain.

Superdomain is uniquely defined in this work as a conserved combination of domains in different proteins. A superdomain, like an individual domain, could occur in different combinations of domains in different proteins. A superdomain could and perhaps often will involve the formation of contacts between its constituent domains in the native protein, similar to the contacts formed between secondary structures in tertiary structure. We illustrate the superdomain concept here, analyze the primary structures of 25 fungal, plant, and animal proteins that contain both a PTP domain and a C2 superdomain in a sequential orientation, and thus test the hypotheses that a superdomain exists and that the PTP-C2 superdomain came into existence prior to the fungi, plant, and animal kingdom divergence, about 1.6 billion years ago.

Results and Discussion

Definition of *superdomain*

Superdomain has previously been employed to describe a proteolytic fragment of endotoxin CryIIIA δ, an inter-domain β-sheet structure in a fragment of human plasma fibronectin, and multi-domain cooperation for predicting protein–protein interactions.[22–24] Residues 280-about 600 of endotoxin CryIIIA δ correspond to the second "domain" of the protein (residues 291–500) and most of the third (residues 501–644).[22] Because the fragment is "stable to…further attack by pepsin," the second and third domains have been described as forming "a cooperative structure, a kind of.'superdomain'." Intact human plasma fibronectin consists of multiple repeats of F_I, $F_{II,}$ and F_{III} modules. In one study, a 2.5-Å resolution structure was obtained of the fragment $6F_I-1F_{II}-2F_{II}-7F_I-8F_I-9F_I$ in the presence of Zn^{2+}.[23] Each module was expected to consist of several module-specific β strands, based on earlier studies. Instead, residues of module $8F_I$ formed two unexpectedly long β strands, which together with residues of modules $7F_I$ and $9F_I$ formed a "superdomain." In Wang *et al.*,[24] *superdomain* signifies domains that "always appear together in individual proteins…[of]…similar biological functions." The authors do not mention the possibility that the constituents of a superdomain could appear individually in proteins, discuss the definition of *superdomain* in an evolutionary context, or suggest a functional role for a superdomain that is not reducible to the sum of the functional properties of the constituent domains. In the present work, by contrast, *superdomain* is given a different and more definite meaning, one that we believe will be more generally useful than in earlier studies. Moreover, the concept is shown to apply to a common structural property of functionally diverse proteins.

The present superdomain concept may be illustrated as follows. Let *A–H* signify eight different conserved and nominally unrelated domains. Polypeptides $A_1B_1C_1D_1E_1$ and $A_2B_2G_2H_2$ comprise two

successive homologous domains, *A* and *B*. If these polypeptides are encoded by different genes, if they are not translations of different splice variants of the same gene, then *AB* is a superdomain on the present definition of the term. *A* and *B* are unrelated, and *AB* is a conserved combination of at least two successive domains which together constitute a heritable unit; *AB* is found in at least two proteins that are not merely different products of a single gene duplication event or of multiple duplications of the same gene. A more compelling case is presented by polypeptide $F_3A_3B_3$, in which superdomain structure is conserved as before but it is located downstream of some other domain. (Further information on the definition of superdomain is provided in Supporting Information.) We argue here that the general case of two successive homologous domains constituting a superdomain applies to the particular case of the PTP and C2 domains in PTEN, tensins and cyclin G Ser/Thr kinase in humans, and to these proteins and others in other species.

Possible advantages of superdomains

Structure conservation is evident across all length scales in the organization of matter in living organisms. The amino acid sequence of the protein actin, for example, is highly conserved across diverse species.[25] Actin monomers will polymerize into filaments under favorable conditions, so both the structure of the monomer and the reversible formation of filaments is conserved. Modular units of protein structure known as domains are conserved.[26] Several examples are noted above. A large percentage of domains are involved in specific binding associated with signaling pathways. SH2 domains and PTB domains, for example, recognize phosphotyrosine or tyrosine in specific binding pockets.[27] These domains are found in many proteins that otherwise differ greatly in structure. Functionality that is both generally useful and achievable by a single domain is conserved in these cases, and binding specificity can often be attributed to amino acid substitutions in the vicinity of the binding pocket. Most individual domains consist of fewer than 150 amino acid residues, limiting the surface area for interaction with a target to < 1500 Å^2 (<9000 Å$^2/6$).[28]

Six possible advantages of superdomain formation and conservation for organisms are noted here. One, tethered domains could increase the interaction surface area with a single receptor molecule and thus increase binding affinity and potency. A possible regulatory advantage is occlusion of a docking site for a competitor protein ligand. Two, tethering domains may increase the complexity of the surface available for binding to a receptor and thus enhance binding specificity. Two domains can form shapes that cannot be formed by a single domain. Three, tethering two or more domains that can recognize different sites on a single receptor simultaneously will increase the double-occupancy of binding sites for fixed concentrations of ligand and receptor by decreasing the overall binding entropy.[29] The odds of double occupancy will increase as the binding of one domain restricts the accessible volume of the other domain; the random walk is done by both domains together rather than by each domain alone.

Four, making a single polypeptide out of two or more domains that can recognize sites on different receptors at the same time could enhance the formation of ternary complexes. If domain *A* binds receptor *X* and domain *B* binds receptor *Y*, for instance, a single protein comprising domains *A* and *B* could enhance the simultaneous formation of *A–X* and *B–Y*, coordinate effects of *A–X* and *B–Y* formation in space and time, and promote the ability of *X* to interact with *Y*. Five, if the binding affinity of one or more tethered modules depends on the phosphorylation state of its receptor, there may be a finer gradation of binding affinity than for separate modules. Superdomain binding affinity could be regulated in this case by kinase and phosphatase activation and inactivation pathways. Six, a combination of domains in a superdomain could enable new levels of control over module function, for example, by way of an allosteric effector. We now demonstrate the existence of the PTP-C2 superdomain.

PTP signature motif

Key regions of the amino acid sequences of this work are aligned in Figure 1. (see Supporting Information Table S1 for the complete sequence alignment.) Substantial structural diversity in the PTP signature motif is evident (panel A). Major features are conserved, but the essential Cys is substituted in about 1/4 of cases (cf. Ref. [16). None of the substitutions is conserved, suggesting either limited selection pressure when Cys is missing or multiple independent instances of persistent loss of activity. The most unusual case is the *Anopheles gambiae* protein (QDREDKHR). Both the nucleophilic Cys and the Gly residue considered crucial for P-loop formation are missing.[30] A sequencing artifact is unlikely, because there is a corresponding sequence change in an *Anopheles aquasalis* protein (JAA99637.1). The *A. gambiae* protein is further distinguished by the absence of an aromatic amino acid about 35 residues upstream of the signature motif. In PTP1B, the corresponding tryptophan side chain may coordinate the substrate in the active site[30]; the distance between the side-chain ζ-carbon atom of the corresponding phenylalanine residue in PTEN and the γ-sulfur atom in the key active-site cysteine residue is 7.4 Å (see Supporting Information Fig. S2 for a hydrophobicity plot). This *A. gambiae* PTP might not even bind a phosphorylated ligand, unlike TNS1 (cf. Ref. [16).

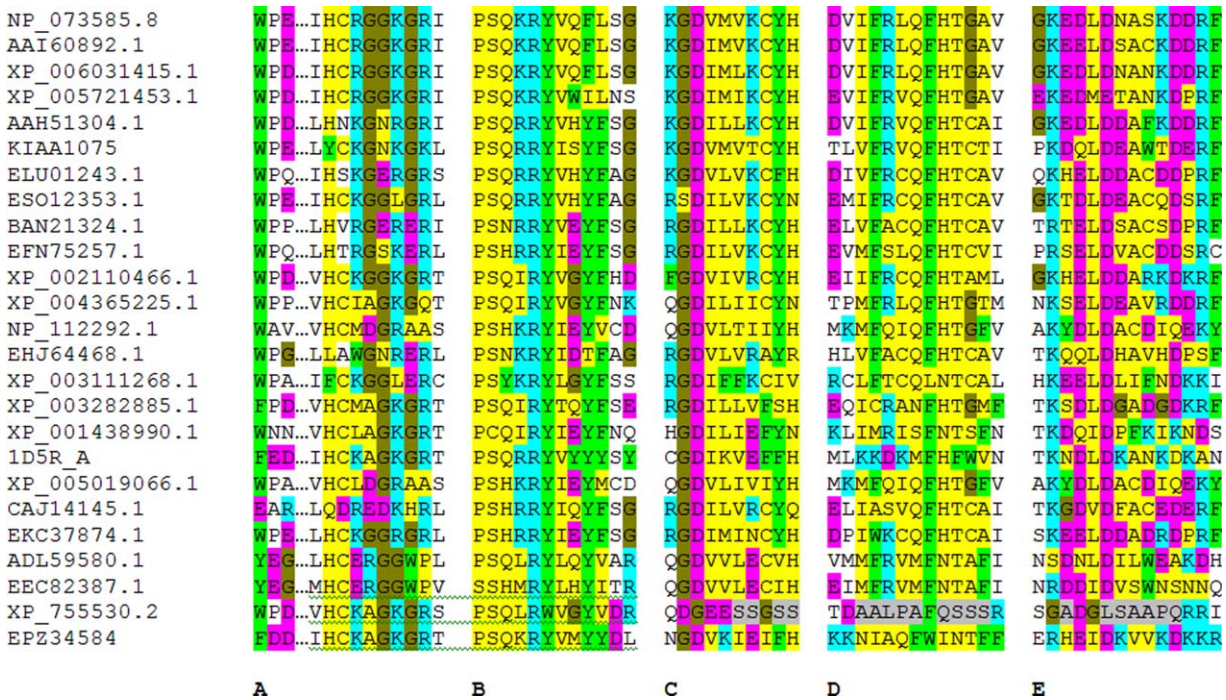

Figure 1. Excerpts of PTP-C2 amino acid sequence alignment. A) Phosphatase signature motif. B) Motif 1, PS(Q/H)(K/R)RYΦXYF. C) Motif 2, Φ_2GDΦ_3(R/K)ΦYH. D) Motif 3, ΦFXΦQFHTΦ_2. E) Motif 4, KX(D/E)L(D/E)X_5(R/K). Green, aromatic residues. Magenta, acidic residues. Cyan, basic residues. Gold, glycine. Yellow, others. Gray, no alignment.

The apparent loss of phosphatase activity but preservation of the PTP-C2 domain organization in disparate proteins suggests that PTP-C2 may have broader significance than phosphoryl group removal or binding. Alternatively, PTP-C2 preservation following loss of activity may be evidence of functional redundancy, possibly owing to gene duplication, in combination with the generally faithful replication of genetic information and the physiological significance of other regions of the same polypeptide. Conservation of phosphatase activity could be unnecessary or disadvantageous in some cases of gene duplication.

Novel conserved sequence motifs in PTP and C2

A second conserved motif in PTP is apparently unique to PTP-C2. PS(Q/H)(K/R)RYΦXYF, Φ indicating "hydrophobic," is identical in human TNS3 and the alligator protein, PSQKRYVQFL, and only modestly diverged in the paramecium protein, PCQIRYIEYF [Fig. 1(B)]. The same motif is identical in the placazoan and *Capsaspora* proteins, PSQIRYVGYF, despite significant sequence divergence elsewhere. The placazoan protein also comprises SH2 and PTB domains, making it TNS-like at the N- and C-termini, and a J-domain is present in the *Capsaspora* protein, making it auxilin- and cyclin G-associated serine/threonine kinase (GAK)-like at the C-terminus. This second conserved motif corresponds to the N-terminal part of a large α helix in PTEN, which forms much of the PTP-C2 domain interface. The conserved tyrosine side chains serve as bridges between the domains, enlarging the surface area of the domain interface.

The PTP-C2 interface in PTEN has a surface area of about 440 $Å^2$, and it is about 70% non-polar (see Supporting Information Table S3). A short linker, just seven residues in PTEN, will make it probable that the domains are docked under usual conditions. The docking probability will presumably increase if hydrophobic side chains in the linker contribute to the domain interface, as does the tyrosine residue in the PTEN linker. Conservation of linker length and hydrophobic character in PTP-C2 in different proteins and across species is evident from the sequence alignment in Supporting Information Table S1.

Conserved motifs are also found in C2 in PTP-C2. One is Φ_2GDΦ_3(R/K)ΦYH [Fig. 1(C)], which forms β strand 10 in PTEN. The conserved glycine residue is in a turn between β strands 9 and 10, and the aspartic acid side chain points at the domain interface. A second motif is ΦFXΦQFHTΦ_2 [Fig. 1(D)]. It forms β strand 11 in PTEN and is located in the domain interface. The second phenylalanine side chain sticks into the core of C2, and histidine side chain is in the interface. A third conserved motif, KX(D/E)L(D/E)X_5(R/K) [Fig. 1(E)], is distinguished by several ionizable side chains. It adopts helical structure at the domain interface in PTEN, forming contacts with the N-terminus of the

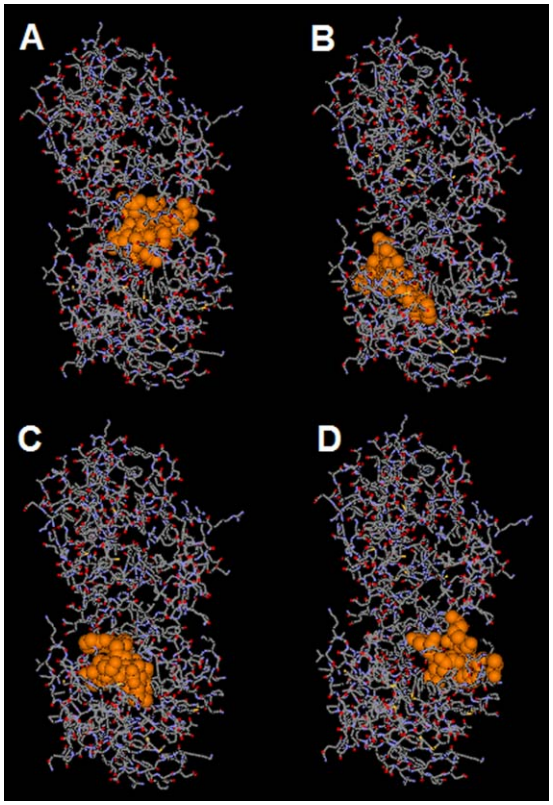

Figure 2. Location in PTEN of the PTP-C2 superdomain conserved motifs. The PTP domain is at the top in each case, the C2 domain at the bottom. A) Motif 1, PS(Q/H)(K/R)RYΦXYF. B) Motif 2, Φ_2GDΦ_3(R/K)ΦYH. C) Motif 3, ΦFXΦQFHTΦ_2. D) Motif 4, KX(D/E)L(D/E)X_5(R/K). All atoms of each residue in each motif are shown space-filled and colored orange. The 1D5R structure was utilized for visualization.

conserved helix in PTP discussed above. The locations in PTEN of the four novel motifs identified here are shown in Figure 2. Each makes a significant contribution to the domain interface. Finally, the sequence data also suggest that β strand-rich C2 is more tolerant of turn-length differences than is mixed α/β PTP in PTP-C2 (see Supporting Information).

Charge properties of PTP-C2

Two further points regarding electric charge are worth noting. One, the p*I* of PTP-C2 is basic for all the animal proteins studied here, regardless of divergence from human TNS3 (circles, Fig. 3). The plant proteins, by contrast, shown as squares, are about 25% identical to human TNS3 in PTP-C2 but are acidic (squares). The physiological significance of these differences is unclear. A distinctive feature of the plant proteins is a formin homology 2 (FH2) domain downstream of PTP-C2. Required for the self-association of formin proteins, FH2 also influences actin polymerization in *Saccharomyces cerevisiae*.[31] In the present animal proteins, by contrast, either no domain is located downstream of PTP-C2, as in PTEN, or PTP-C2 is followed by J, SH2 or PTB, as in GAK and tensins.[12] PTP-C2 in the aquatic organisms (triangles) is less basic than in the terrestrial animals. The arrow indicates a possible exception. PTEN-like, this *Helobdella robusta* protein consists of just PTP-C2. The p*I* of human PTEN is 8.05 (star).

Two, analysis of the signature motif [Fig. 1(A)] suggests the importance of an acidic side chain in the active site. In human TNS3, for example, the sequence is WP*E*...IH*C*RGGKGRI. The Glu side chain, though distant in sequence from the Cys nucleophile, may function as a general acid in the phosphatase reaction mechanism.[30] This residue is Asp in human PTEN. In about 1/2 of the present proteins, by contrast, the corresponding side chain cannot ionize. In the *Capitella teleta* protein this residue is Gln, and in the *Riptortus pedestris* protein it is Pro. Ten of 12 such cases are correlated with the in-substitution of an acidic side chain in the signature motif. In the *C. teleta* protein the motif is WPQ...IH*S*KG*E*RGRS. The *Capsaspora owczarzaki* and *Paramecium tetraurelia* proteins are exceptions.

PTP-C2 evolution

Additional evidence supports the claimed existence of a PTP-C2 superdomain, that is, the inheritance of the two domains a single structural unit. Figure 4 shows a schematic of the molecular architecture of exemplars of the present set of proteins. A key example not shown is the human putative membrane protein EAX08222, in which PTP-C2 is at the

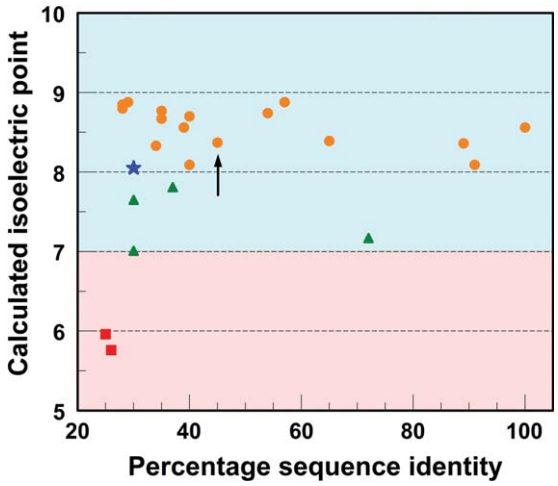

Figure 3. Calculated isoelectric point versus nominal percentage identity for the present PTP-C2 superdomain sequences. The comparisons were made with respect to human TNS3. A cyan background represents basicity, and a rose background, acidity. Data points for the two plant proteins are shown as red squares, aquatic organisms as green triangles, and human PTEN as a blue star. An arrow highlights the leech protein.

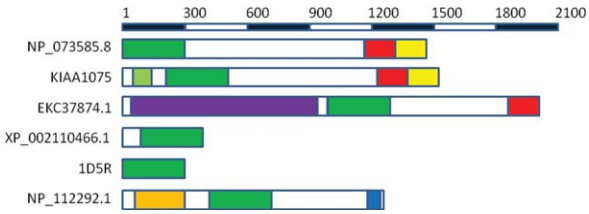

Figure 4. Molecular architecture of PTP-C2 superdomain proteins. Domain composition is diverse. Green, PTP-C2. Red, SH2. Yellow, PTB. Olive, cysteine-rich Zn^{2+}-binding. Violet, karyopherin β, comprising armadillo/β-catenin-like repeats. Orange, Ser/Thr kinase. Blue, DNA J. The scale bar represents amino acid residues.

C-terminus.[12] In short, PTP-C2 occurs in diverse locations in different proteins, the domain compositions of these proteins are rather heterogeneous, and the known or probable functions of these proteins are diverse.

The TNS-like PTP domains, TNS-like C2 domains, and PTP-C2 superdomains of this work have been compared at a more detailed level of structure. Figure 5 presents phylogenetic trees involving all of the present proteins, each based on human TNS3 (NP_073585.8), or tumor endothelial marker 6.[32] The trees will now be discussed in turn.

For PTP, NP_073585.8 and AAI60892.1 (rat TNS3) are grouped together, as are AAH51304.1 (human TNS 1) and KIAA1075 (human TNS2), located nearby. ELU01243.1 and ESO12353.1, PTEN-like proteins in an annelid worm and a leech, are relatively close to EKC37874.1, Pacific oyster importin subunit β-1. All these PTP domains are within a few residues of the apparent N-terminus of the protein. The next several proteins—XP_003111268.1, a nematode protein, CAJ14145.1, a mosquito protein, BAN21324.1, a bean bug protein, EFN75257.1, an ant protein, and EHJ64468.1, a monarch butterfly protein—form a group. The PTP domain is at least 39 residues from the apparent N-terminus in members of the group, and the ant protein, for example, consists of just PTP-C2, like PTEN (1D5R). The next major group includes XP_002110466.1, a *Trichoplax adhaerens* protein that, like human tensins 1–4, comprises a SH2 domain and a PTB domain at the C-terminus. *T. adhaerens* is the simplest multicellular organism known. The PTP domain of 1D5R is closest to EPZ34584 and XP_755530.2, both fungal proteins. Also in this group are XP_003282885.1, from a *Dictyostelium*, and XP_001438990.1 from a *Paramecium*. XP_004365225.1, a *Capsaspora* GAK/auxilin-like protein, NP_112292.1, rat GAK, and XP_005019066.1, mallard duck GAK, form a separate cluster. There is Ser/Thr kinase domain upstream of PTP in these proteins. In ADL59580.1 and EEC82387.1, both plant proteins, PTP is close to the N-terminus, as in human TNS1 and TNS3, but others rather distant from the others analyzed here.

The C2 comparison is both similar to and quite different from that for PTP alone. The main difference now is that, whereas the PTP sequences could be sorted into several relatively large groups, the C2 sequences could not be. This is apparently because sequence conservation is, on the whole, greater in TNS-like PTP domains than in TNS-like C2 domains. The two cases are similar in that the sequence sets are ordered in roughly the same way. Moreover, the GAK sequences are clustered together as before, as are the plant proteins. A key difference from the PTP comparison is that human PTEN is

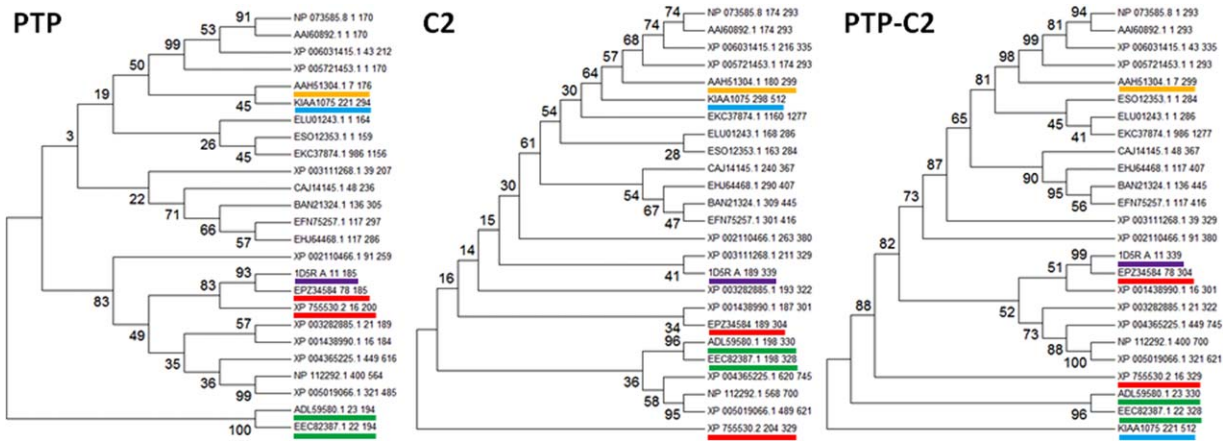

Figure 5. PTP-C2 superdomain phylogenetic trees. All sequences of this work were included in comparisons to human TNS3 PTP, C2 or PTP-C2. Human TNS1, orange. Human TNS2, cyan. Human PTEN, violet. The two fungal proteins, red. The two plant proteins, green. Each comparison is quantified by a percentage of the bootstrap consensus tree. The higher the number, the higher the reliability of the branch. See Materials and Methods section for the list of proteins analyzed and a description of the procedure. Supporting Information Table S2 presents details of the proteins included in the comparison. The text of Supporting Information contains further details of the comparison.

not closest to the fungal proteins. Instead, PTEN C2 is closest to that of a nematode protein. This too can probably be attributed to a difference in sequence conservation between domains and, presumably, a difference in function conservation. TNS-like C2 sequences are particularly diverged near the C-terminus, which in human PTEN is the part of the C2 domain that is farthest away from the domain interface.

The PTP-C2 tree is consistent with results for PTP and C2. The plant proteins form a distinct group. Human PTEN is clustered with the GAKs but closest to a fungal protein (EPZ34584). A remarkable similarity of PTEN and EPZ34584 is 100% amino acid sequence identity in active site residues: IHCKAGKGRTGVMIC. Another surprise is that KIAA1075 (human TNS2) is so distant from human TNS3; human TNS1 (AAH51304.1) is much closer to TNS3. This outcome is consistent with a hypothesis of TNS gene origin.[12] On this view, tensin proto-gene formation involved preformed PTP-C2. Duplication of the tensin proto-gene yielded a proto-tensin 1/3 gene and a proto-tensin 2/4 gene. Both of these genes duplicated, giving separate genes for tensin 1–4. Then, at uncertain points later in time, TNS 2 acquired a cysteine-rich Zn^{2+}-binding domain at the N-terminus and TNS 4 lost PTP-C2.

Finally, it should be clear that the three trees are not adduced as evidence in support of the claimed existence of a PTP-C2 superdomain. Instead, they help to visualize the divergence of PTP-C2. They also suggest, however, that divergence has been domain-dependent, despite tethering. The structures and functions of these domains will be different, even if their structures and functions are integrated. Presumably, then, domain-specific selection pressures are different at one or more levels between gene structure and protein structure. It seems possible, for example, that point mutations have been better tolerated in one domain than the other, and that gene structure is more stable in one domain than the other.

PTEN loss of function mutations

Clues concerning the parallel inheritance of TNS-like PTPs and C2s come from further analysis of human PTEN. Exon 6 encodes residues from before the final PTP helix, PSQRRYVYYY (helix 5, residues 169–178), to well into C2.[33] This conserved motif [Fig. 2(A)], and the noted conserved motifs in C2 [Fig. 2(B–D)], form the domain interface. Uncompensated changes of shape or charge complementarity in the interface could reduce the thermostability of PTP-C2, PTP or C2 and thus result in loss of function (e.g., Ref. [34). Human PTEN variants are of considerable medical interest.[33] Mutations are known to have occurred in the novel motifs identified in the present work. S170N (variant 026266), for example, abrogates activity toward phosphoinositides but not phospholipid membrane association, despite being outside the active site.[35] This Ser side chain points directly into the domain interface. Similarly, S170R (variant 007470) in Bannayan-Riley-Ruvalcaba syndrome,[36] a disease marked by macrocephaly, noncancerous tumors and tumor-like growths,[37] severely reduces activity toward proteins and eliminates it toward inositides.[35] Q171A/E results in a 75% reduction in activity. The Gln side chain points into C2. R173C (variant 026267) in endometrial hyperplasia displays no activity toward inositides but bind phospholipid membranes.[35,38] The Arg side chain points toward C2, the guanidinium group remaining solvent-exposed. R173H (variant 026268), R173P (variant 026269), and Y174N (variant 026270) show no activity toward inositides.[35] The Tyr side chain points directly into C2.

Conclusions

Superdomain has been uniquely defined as a conserved combination of different globular domains in different proteins. A superdomain is a step beyond consecutive modular folding units in protein organization. A superdomain thus represents a level of the protein structure hierarchy that has not been identified before now. A superdomain might represent a specialized structure or function that is too complex for encoding in a single domain. For instance, regulation of protein function might involve an allosteric mechanism that depends on interactions between the modular units of a superdomain, or cellular processes might be inefficiently realized when the modular units are encoded as separate polypeptides. The identification of superdomains could advance knowledge of the relationship of archaebacteria, bacteria and eukaryota, and the relationship of fungi, plantae, and animalia, and it could provide insight on the molecular basis of cell function.

The present analysis provides compelling support for the hypothesis that TNS-like PTP-TNS-like C2 constitutes a superdomain on the present definition. PTP-C2 is the first superdomain identified. PTP-C2 came into existence prior to the divergence of eukaryotes, before 1 but apparently after 2 billion years ago,[39,40] possibly by the fusion of two preexisting genes. PTP-C2 is apparently inessential for life, but it may be essential in eukaryotes or fungi. Amino acid sequence comparisons suggest that loss of phosphatase activity in TNS-like PTP is better tolerated by organisms than loss of the structural integrity of PTP-C2. The interdependence of TNS-like PTP and TNS-like C2 implied by superdomain formation may have structural and functional aspects. For example, the interface could make a substantial contribution to the thermostability of PTP, C2 or both domains, and thus influence

functionality. In any case, TNS-like PTP and TNS-like C2 interdependence is corroborated by the demonstrated conservation of amino acid sequence in the domain interface and the seriousness of interface-related mutations in human PTEN.

Materials and Methods

Amino acid sequence data were obtained from the non-redundant protein sequence database of National Center for Biotechnology Information in December 2013–November 2014, and we compared by protein–protein BLAST. We also accessed the *Arabidopsis thaliana*, *Oryza sativa,* and microbes databases available at NCBI then. Sequence comparisons with the BLOSUM62 matrix yielded nominal values of sequence identity. Minor adjustments were made to obtain the PTP-C2 alignment shown in Supporting Information Table S1.

The accession codes of the polypeptide sequences of the present data set are the following: NP_073585.8 (3), AAI60892.1 (3), XP_006031415.1 (45), XP_005721453.1 (3), AAH51304.1 (9), KIAA1075 (223), ELU01243.1 (1), ESO12353.1 (1), BAN21324.1 (138), EFN75257.1 (119), XP_002110466.1 (93), XP_004365225.1 (451), NP_112292.1 (402), EHJ64468.1 (119), XP_003111268.1 (41), XP_003282885.1 (23), XP_001438990.1 (18), 1D5R (also a Protein Data Bank accession code, 13), XP_005019066.1 (323), CAJ14145.1 (50), EKC37874.1 (988), ADL59580.1 (25), EEC82387.1 (24), XP_755530.2 (18), and EPZ34584 (19), where in each case the nominal first residue of the PTP domain is indicated in parentheses. The proteins analyzed include at least one tensin, one PTEN and one GAK/auxilin, none of which contains β-catenin-like repeats, and at least one protein that contains β-catenin-like repeats; the selected proteins constitute a structurally and functionally diverse set. Further information is provided in Supporting Information: Table S1 shows all PTP-C2 sequence alignment data for the selected proteins, not only the conserved regions highlighted in Figure 1, Supporting Information Table S2 provides background information on protein function, if known, and genome, Supporting Information Table S3 presents surface area calculations, and Supporting Information Table S4 presents isoelectric point information.

The crystallographic structure of PTEN was obtained from the PDB (accession code 1D5R). Calculated values of p*I* were obtained in December 2013–November 2014 with the online ExPASy Compute pI/Mw tool. The input in each case was the amino acid sequence shown in Supporting Information Table S1. The phylogenetic tree was obtained as follows. The curated PTP-C2 sequence alignment was imported into MEGA5, software for molecular evolutionary genetics analysis.[41] Maximum likelihood estimation was then used to build the phylogenetic tree by bootstrapping 100 times.

References

1. Linderstrøm-Lang KU (1952) Proteins and enzymes. Lane Medical Lectures, Stanford University Publications, University Series, Medical Sciences, Vol. VI. California: Stanford University Press.
2. Antonini E (1965) Interrelationship between structure and function in hemoglobin and myoglobin. Physiol Rev 45:123–170.
3. McKenzie HA, White FH (1991) Lysozyme and α-lactalbumin: structure, function, and interrelationships. Adv Prot Chem 41:173–315.
4. Orengo CA, Jones DT, Thornton JM (1994) Protein superfamilies and domain superfolds. Nature 372:631–634.
5. Tonks NK (2006) Protein tyrosine phosphatases: from genes, to function, to disease. Nat Rev Mol Cell Biol 7: 833–846.
6. Nalefski EA, Falke JJ (1996) The C2 domain calcium-binding motif: structural and functional diversity. Protein Sci 5:2375–2390.
7. Nagano N, Orengo CA, Thornton JM (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol 321:741–765.
8. Apic G, Gough J, Teichmann SA (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol 310:311–325.
9. Potts JR, Campbell ID (1996) Structure and function of fibronectin modules. Matrix Biol 15:313–320.
10. Filippakopoulos P, Müller S, Knapp S (2009) SH2 domains: modulators of nonreceptor tyrosine kinase activity. Curr Opin Struct Biol 19:643–649.
11. Uhlik MT, Temple B, Bencharit S, Kimple AJ, Siderovski DP, Johnson GL (2005) Structural and evolutionary division of phosphotyrosine binding (PTB) domains. J Mol Biol 345:1–20.
12. Haynie DT (2014) Molecular physiology of the tensin brotherhood of integrin adaptor proteins. Proteins Struct Funct Bioinformat 82:1113–1127.
13. Lee J-O, Yang H, Georgescu M-M, Di Cristofano A, Maehama T, Shi Y, Dixon JE, Pandolfi P, Pavletich NP (1999) Crystal structure of the PTEN tumor suppressor: implications for its phosphoinositide phosphatase activity and membrane association. Cell 99:323–334.
14. Patterson KI, Brummer T, O'Brien PM, Daly RJ (2009) Dual-specificity phosphatases: critical regulators with diverse cellular targets. Biochem J 418:475–489.
15. Yuvaniyama J, Denu JM, Dixon JE, Saper MA (1996) Crystal structure of the dual specificity protein phosphatase VHR. Science 272:1328–1331.
16. Haynie DT, Ponting CP (1996) The N-terminal domains of tensin and auxilin are phosphatase homologues. Protein Sci 5:2643–2646.
17. Li J, Yen C, Liaw D, Podsypanina K, Bose S, Wang SI, Puc J, Miliaresis C, Rodgers L, McCombie R, Bigner SH, Giovanella BC, Ittmann M, Tycko B, Hibshoosh H, Wigler MH, Parsons R (1997) *PTEN*, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. Science 275:1943–1947.
18. Steck PA, Pershouse MA, Jasser SA, Yung WKA, Lin H, Ligon AH, Langford LA, Baumgard ML, Hattier T, Davis T, Frye C, Hu R, Swedlund B, Teng DHR, Tavtigian SV (1997) Identification of a candidate tumour suppressor gene, MMAC1, at chromosome 10q23.3 that is mutated in multiple advanced cancers. Nat Genet 15:356–362.
19. Maehama T, Dixon JE (1998) The tumor suppressor, PTEN/MMAC1, dephosphorylates the lipid second messenger, phosphatidylinositol 3,4,5-trisphosphate. J Biol Chem 273:13375–13378.

20. Benes CH, Wu N, Elia AE, Dharia T, Cantley LC, Soltoff SP (2005) The C2 domain of PKCdelta is a phosphotyrosine binding domain. Cell 121:271–280.
21. Zhang D, Aravind L (2010). Identification of novel families and classification of the C2 domain superfamily elucidate the origin and evolution of membrane targeting activities in eukaryotes. Gene 469: 18–30.
22. Ort P, Zalunin IA, Gasparov VS, Chestukhina GG, Stepanov VM (1995) Domain organization of *Bacillus thuringiensis* CryIIIA delta-endotoxin studied by denaturation in guanidine hydrochloride solutions and limited proteolysis. J Prot Chem 14:241–249.
23. Graille M, Pagano M, Rose T, Ravaux MR, van Tilbeurgh H (2010) Zinc induces structural reorganization of gelatin binding domain from human fibronectin and affects collagen binding. Structure 18:710–718.
24. Wang R-S, Wang Y, Wu L-Y, Zhang X-S, Chen L (2007) Analysis on multi-domain cooperation for predicting protein-protein interactions. BMC Bioinformat 8:391.
25. Dominguez R, Holmes KC (2011) Actin structure and function. Annu Rev Biophys 40:169–186.
26. Schultz J, Milpetz F, Bork P, Ponting CP (1998) SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci USA 95:5857–5864.
27. Shoelson SE (1997) SH2 and PTB domain interactions in tyrosine kinase signal transduction. Curr Opin Chem Biol 1:227–234.
28. Jainin J, Bahadur RP, Chakrabarti P (2008) Protein-protein interaction and quaternary structure. Q Rev Biophys 41:133–180.
29. Weber G (1975) Energetics of ligand binding to proteins. Adv Prot Chem 29:1–83.
30. Jia Z, Barford D, Flint AJ, Tonks NK (1995) Structural basis for phosphotyrosine peptide recognition by protein tyrosine phosphatase 1B. Science 268:1754–1758.
31. Goode BL, Eck MJ (2007) Mechanism and function of formins in the control of actin assembly. Annu Rev Biochem 76:593–627.
32. St Croix B, Rago C, Velculescu V, Traverso G, Romans KE, Montgomery E, Lal A, Riggins GJ, Lengauer C, Vogelstein B, Kinzler KW (2000) Genes expressed in human tumor endothelium. Science 289:1197–1202.
33. Bonneau D, Longy M (2000) Mutations of the human PTEN gene. Hum Mutat 16:109–122.
34. Haynie DT (2008) Biological thermodynamics. 2 ed. Cambridge: Cambridge University Press, 6 p.
35. Han SY, Kato H, Kato S, Suzuki T, Shibata H, Ishii S, Shiiba K, Masuno S, Kanumaru R, Ishioka C (2000) Functional evaluation of PTEN missense mutations using in vitro phosphoinositide phosphatase assay. Cancer Res 60:3147–3151.
36. Eng C (2003) PTEN: one gene, many syndromes. Hum Mutat 22:183–198.
37. Lynch NE, Lynch SA, McMenamin J, Webb D (2009) Bannayan-Riley-Ruvalcaba syndrome: a cause of extreme macrocephaly and neurodevelopmental delay. Arch Dis Child 94:553–554.
38. Maxwell GL, Risinger JI, Gumbs C, Shaw H, Bentley RC, Barrett JC, Berchuck A, Futreal PA (1998) Mutation of the PTEN tumor suppressor in endometrial hyperplasia. Cancer Res 58:2500–2503.
39. Doolittle RF, Feng D-F, Tsang S, Cho G, Little E (1996) Determining divergence times of the major kingdoms of living organisms with a protein clock. Science 271: 470–477.
40. Wang DY, Kumar S, Hedges SB (1999) Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. Proc Biol Sci 266:163–171.
41. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739.